On the power of adaptivity in statistical adversaries

Guy Blanc GBLANC@STANFORD.EDU
Stanford University

Jane Lange JLANGE@MIT.EDU
Massachusetts Institute of Technology

Ali Malik MALIKALI@CS.STANFORD.EDU
Stanford University

Li-Yang Tan LIYANG@CS.STANFORD.EDU
Stanford University

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We initiate the study of a fundamental question concerning adversarial noise models in statistical problems where the algorithm receives i.i.d. draws from a distribution $\mathcal{D}$. The definitions of these adversaries specify the *type* of allowable corruptions (noise model) as well as *when* these corruptions can be made (adaptivity); the latter differentiates between oblivious adversaries that can only corrupt the distribution $\mathcal{D}$ and adaptive adversaries that can have their corruptions depend on the specific sample S that is drawn from $\mathcal{D}$.

We investigate whether oblivious adversaries are effectively equivalent to adaptive adversaries, across all noise models studied in the literature, under a unifying framework that we introduce. Specifically, can the behavior of an algorithm $\mathcal{A}$ in the presence of oblivious adversaries always be well-approximated by that of an algorithm $\mathcal{A}'$ in the presence of adaptive adversaries? Our first result shows that this is indeed the case for the broad class of *statistical query* algorithms, under all reasonable noise models. We then show that in the specific case of *additive noise*, this equivalence holds for *all* algorithms. Finally, we map out an approach towards proving this statement in its fullest generality, for all algorithms and under all reasonable noise models.

Keywords: Statistical problems, Adversary models, Robustness, Statistical query algorithms

1. Introduction

The possibility of noise pervades most problems in statistical estimation and learning. In this paper we will be concerned with *adversarial* noise models, as opposed to the class of more benign random noise models. Adversarial noise models are the subject of intensive study across statistics (Huber, 1964; Hampel, 1971; Tukey, 1975), learning theory (Valiant, 1985; Haussler, 1992; Kearns and Li, 1993; Kearns et al., 1994; Bshouty et al., 2002), and algorithms (Diakonikolas et al., 2019; Lai et al., 2016; Charikar et al., 2017; Diakonikolas and Kane, 2019). The definition of each model specifies:

1. The *type* of corruptions allowed. For example, the adversary may be allowed to add arbitrary points (additive noise (Huber, 1964; Valiant, 1985)), or in the context of supervised learning, allowed to change the labels in the data (agnostic noise (Haussler, 1992; Kearns et al., 1994)).

2. The *adaptivity* of the adversary.

The latter is the focus of our work. Consider any statistical problem where the algorithm is given i.i.d. draws from a distribution $\mathcal{D}$. On one hand we have *oblivious* adversaries: such an adversary corrupts $\mathcal{D}$ to a different distribution $\widehat{\mathcal{D}}$, from which the algorithm then receives a sample. On the other hand we have *adaptive* adversaries: such an adversary first draws a sample $\boldsymbol{S}$ from $\mathcal{D}$, and upon seeing the specific outcomes in $\boldsymbol{S}$, corrupts it to $\widehat{\boldsymbol{S}}$ which is then passed on to the algorithm. One can further consider adversaries with intermediate adaptive power, but we think of this as a dichotomy for now. A coupling argument shows that adaptive adversaries are at least as powerful as oblivious ones (Diakonikolas et al., 2019; Zhu et al., 2019). In this work we investigate whether they can be *strictly* more powerful.

Question 1 *Fix the type of corruptions allowed. Is it true that for any algorithm $\mathcal{A}$, there is an algorithm $\mathcal{A}'$ whose behavior in the presence of adaptive adversaries well-approximates that of $\mathcal{A}$ in the presence of oblivious adversaries?*

The distinction between oblivious and adaptive adversaries is frequently touched upon in works concerning statistical problems. Sometimes this distinction is brought up in service of emphasizing that the algorithms given in these works are robust against adaptive adversaries; other times it is brought up when the algorithms are shown to be robust against oblivious adversaries, and the viewpoint of an adaptive corruption process is provided as intuition for the noise model. However, the relative power of oblivious and adaptive adversaries in the statistical setting has not been systematically considered in the literature.

1.1. Our contributions

1.1.1. A UNIFIED FRAMEWORK FOR CHARACTERIZING DATA ADVERSARIES.

To reason generally about Question 1, we associate every type of allowable corruptions with a cost function ρ between distributions. An oblivious ρ-adversary therefore corrupts $\mathcal{D}$ to some $\widehat{\mathcal{D}}$ that is η-close with respect to ρ, meaning that $\rho(\mathcal{D}, \widehat{\mathcal{D}}) \leq \eta$. An adaptive ρ-adversary corrupts a sample S drawn from $\mathcal{D}$ to some $\widehat{S}$ such that the uniform distribution over $\widehat{S}$ is η-close with respect to ρ to that over S. For example, when $\rho(\mathcal{D}, \widehat{\mathcal{D}})$ is the total variation distance between $\mathcal{D}$ and $\widehat{\mathcal{D}}$, the resulting adaptive adversary represents *nasty noise* as defined in Bshouty et al. (2002), where the adversary is allowed to change an arbitrary η-fraction of the points in S. We discuss this framework in more detail in Section 3.

1.1.2. YES TO QUESTION 1 FOR ALL SQ ALGORITHMS

Our first result is an affirmative answer to Question 1 for the broad class of *statistical query* (SQ) algorithms. Our proof will require a mild assumption on this cost function, the precise statement of which we defer to the body of the paper. For now, we simply refer to cost functions satisfying this assumption as "reasonable", and mention that it is easily satisfied by all standard noise models, and can be seen to be necessary for our result to hold.

Theorem 1 (SQ algorithms are robust to adaptive adversaries) [1] *For all reasonable cost functions ρ and SQ algorithms $\mathcal{A}$, the behavior of $\mathcal{A}' := \mathcal{A}$ in the presence of adaptive ρ-adversaries well-approximates that of $\mathcal{A}$ in the presence of oblivious adversaries.*

1. See Theorem 5 for the formal version of this theorem.

A key ingredient in our proof of Theorem 1 is a novel reduction, using duality, from a k-query SQ algorithm to a single "representative" SQ.

In other words, the SQ framework neutralizes adaptive adversaries into oblivious ones, and in the context of Question 1, we can take $\mathcal{A}'$ to be $\mathcal{A}$ itself. Looking ahead to our other results, we remark that such a statement cannot be true for all algorithms: there are trivial examples of (non-SQ) algorithms $\mathcal{A}$ for which $\mathcal{A}'$ has to be a modified version of $\mathcal{A}$.[2]

Theorem 1 adds to the already-deep connections between the SQ framework and noise tolerance. The SQ framework was originally introduced in learning theory, where it continues to be influential in the design of learning algorithms that are resilient to *random classification noise* (Kearns, 1998). Most relevant to the topic of this paper, to our knowledge across all noise models, all existing statistical algorithms that have been shown to be robust to adversarial noise can be cast in the SQ framework.

1.1.3. YES TO QUESTION 1 FOR ADDITIVE NOISE

Our second result is an affirmative answer to Question 1 for one of the most natural types of corruptions:

Theorem 2 [3] *The answer to Question 1 is "yes" for additive noise.*

In additive noise, an oblivious adversary corrupts $\mathcal{D}$ to $\widehat{\mathcal{D}} = (1-\eta)\mathcal{D} + \eta\mathcal{E}$ for an arbitrary distribution $\mathcal{E}$ of their choosing. An adaptive adversary, on the other hand, gets to inspect the sample S drawn from $\mathcal{D}$, and adds $\frac{\eta}{1-\eta}|S|$ many arbitrary points of their choosing to S. The oblivious version of additive noise was introduced by Huber (Huber, 1964) and has become known as Huber's contamination model; the adaptive version is commonly called "data poisoning" in the security and machine learning literature.

Additive noise also captures the well-studied *malicious noise* (Definition 16) from learning theory (Valiant, 1985) (see also (Kearns and Li, 1993)). Theorem 2 therefore shows that Huber's contamination model, the malicious noise model, and the adaptive version of additive noise, are in fact all equivalent. Our proof of Theorem 2 is constructive: we give an explicit description of how $\mathcal{A}'$ can be obtained from $\mathcal{A}$, and $\mathcal{A}'$ preserves the computational and sample efficiency of $\mathcal{A}$ up to polynomial factors.

1.1.4. YES TO QUESTION 1 IN ITS FULLEST GENERALITY? AN APPROACH VIA SUBSAMPLING

Our proof of Theorem 2 is actually an instantiation of a broader approach towards answering Question 1 affirmatively in its fullest generality: showing that the answer is "yes" for all (reasonable) types of allowable corruptions and all algorithms. We introduce the following definition:

Definition 1 (Neutralizing filter) *Let ρ be a cost function and $\mathcal{A}$ be an algorithm for a statistical problem over a domain $\mathcal{X}$. We say that a randomized function $\Phi : \mathcal{X}^* \to \mathcal{X}^*$ is a neutralizing filter for $\mathcal{A}$ with respect to ρ if the following holds. For all distributions $\mathcal{D}$, with high probability over the draw of $\boldsymbol{S}$ from $\mathcal{D}$, the behavior of*

2. One such example is $\mathcal{D} = \text{Bernoulli}(\frac{1}{2})$ and $\mathcal{A} = \mathbb{1}[\text{number of 1's in the sample is 0 mod 100}]$. As the sample size grows, an oblivious adversary can barely change the acceptance probability of $\mathcal{A}$, whereas an adaptive one can change it completely.
3. See Theorem 7 for the formal version

$\mathcal{A}$ on $\Phi(\widehat{\boldsymbol{S}})$, where $\widehat{\boldsymbol{S}}$ is a corruption of $\boldsymbol{S}$ by an adaptive ρ-adversary,

well-approximates the behavior of

$\mathcal{A}$ on a sample from $\widehat{\mathcal{D}}$, where $\widehat{\mathcal{D}}$ is a corruption of $\mathcal{D}$ by an oblivious ρ-adversary.

Perhaps the most natural filter in this context is the *subsampling* filter. For an n-sample algorithm $\mathcal{A}$, we request for a larger sample of size $m \geq n$, allow the adaptive adversary to corrupt it, and then run $\mathcal{A}$ on a size-n subsample of the corrupted size-m sample. We call this the "$m \to n$ subsampling filter" and denote it as $\Phi_{m\to n}$. The hope here is for the randomness of the subsampling step to neutralize the adaptivity of the adversary.

The efficiency of the subsampling filter is measured by the overhead in sample complexity that it incurs, i.e. how much larger m is relative to n. Subsampling from a sample that is roughly the size of the domain of course makes adaptive adversaries equivalent to oblivious ones, but this renders a sample-efficient algorithm inefficient. We are interested whether the subsampling filter can be effective while only incurring a mild overhead in sample complexity.

We propose the following conjecture as a general approach towards answering Question 1:

Conjecture 1 *For all reasonable cost functions ρ and n-sample algorithms $\mathcal{A}$, the subsampling filter $\Phi_{m\to n}$ is a neutralizing filter for $\mathcal{A}$ with respect to ρ with $m = \mathrm{poly}(n, \log(|\mathcal{X}|))$.*

We obtain Theorem 2 by proving Conjecture 1 in the case of ρ being additive noise. While we have not been able to prove Conjecture 1 for all ρ's and all algorithms, we can show the following:

Theorem 3 (If it is possible, subsampling neutralizes adaptivity) [4] *Let ρ be a cost function and $\mathcal{A}$ be an n-sample algorithm. Suppose there is an m-sample algorithm $\mathcal{A}'$ whose behavior in the presence of adaptive ρ-adversaries well-approximates that of $\mathcal{A}$ in the presence of oblivious ρ-adversaries. Then $\Phi_{M\to n}$ is a neutralizing filter for $\mathcal{A}$ with respect to ρ with $M = O(m^2)$.*

We note that in the context of Theorem 3, we do not require $\mathcal{A}'$ to be computationally efficient: as long as $\mathcal{A}'$ is sample efficient, then our resulting algorithm, the subsampling filter applied to $\mathcal{A}$, inherits the computational efficiency of $\mathcal{A}$.

Finally, we show that the bound on m in Conjecture 1 cannot be further strengthened to be independent of $|\mathcal{X}|$, the size of the domain:

Theorem 4 (Subsampling lower bound) [5] *Let ρ be the cost function for additive noise and η be the corruption budget. There is an n-sample algorithm $\mathcal{A}$ such that for $m \leq O_\eta(n \log(|\mathcal{X}|)/\log^2 n)$, $\Phi_{m\to n}$ is not a neutralizing filter for $\mathcal{A}$ with respect to ρ.*

While Theorem 4 shows that some dependence on $|\mathcal{X}|$ is necessary, we remark that the $\log(|\mathcal{X}|)$ dependence in Conjecture 1 is fairly mild—this is the description length of a sample point $x \in \mathcal{X}$. Theorem 4 also shows that the quantiative bounds that we achieve for Theorem 2 has an optimal dependence on $|\mathcal{X}|$.

4. See Theorem 6 for the formal version
5. See Theorem 8 for the formal version

1.2. Other related work

A separation result Recently, Deng et al. gave a separation between the adaptive adversary, which they call "data-aware", and the oblivious adversary (Deng et al., 2021). Specifically they showed that there are settings in which a natural Lasso-based algorithm for feature selection will often fail to select the correct features in the presence of an adaptive additive adversary, but would succeed in the presence of an oblivious additive adversary. Our Theorem 2 implies that this Lasso-based algorithm would succeed if it used the subsampling filter to preprocess its sample.

The online and dynamic setting While the focus of our work is on the statistical setting, the distinction between adaptive and oblivious adversaries has also been the subject of recent study in the *online* (Haghtalab et al., 2021; Alon et al., 2021) and *dynamic* (Beimel et al., 2022) setting, albeit with a notably different notion of adaptivity. In these settings, the adaptive adversaries can change the input distribution in response to the previous behavior of the algorithm, while oblivious adversaries must choose a fixed input distribution before the algorithms run.

Adaptive data analysis We emphasize the distinction between the focus of our work and the recent fruitful line of work on adaptive data analysis (Dwork et al. (2015); Hardt and Ullman (2014); Steinke and Ullman (2015); Bassily et al. (2021)). The focus of our work is on the adaptivity of the *adversary*, whereas the focus of this line of work is on the adaptivity of the *SQ algorithm*. Throughout this work, we reserve the use of "adaptive" to refer to the adversary, and all SQ algorithms will inherently be adaptive.

1.3. Discussion and future work

Implications of our results Theorem 1 says that for all reasonable cost functions, all SQ algorithms (existing and future ones) that are resilient to oblivious adversaries are "automatically" also resilient to adaptive adversaries. Likewise, lower bounds against adaptive adversaries immediately yield lower bounds against oblivious ones. The same remark further applies for all algorithms in the case of additive noise and malicious noise, by Theorem 2.

As a concrete example, we recall that the agnostic learning framework was originally defined with respect to oblivious adversaries (Haussler, 1992; Kearns et al., 1994). As in the PAC model there a concept class $\mathcal{C}$, but the target function f is no longer assumed to lie within $\mathcal{C}$—hence the name of the model. The learning algorithm is expected to achieve error close to opt, the distance from f to $\mathcal{C}$. However, many papers on agnostic learning provide the viewpoint of an adaptive corruption process as intuition for the model: the data *is* assumed to be a labeled according to a function $f \in \mathcal{C}$, but an adversary corrupts an opt fraction of the labels given to the learning algorithm. This adaptive version was subsequently defined as a separate model called *nasty classification noise* (Bshouty et al., 2002) (as a special case of the nasty sample noise model introduced in that paper). Theorem 1 therefore shows that the agnostic learning model and the nasty classification noise model are in fact equivalent when it comes to SQ algorithms.

When can quantifiers be swapped? The distinction between oblivious and adaptive adversaries can be viewed as a difference in the order of "for all" and "with high probability" quantifiers in the performance guarantees of statistical algorithms. An algorithm succeeds in the presence of oblivious adversaries if *for all* distributions $\widehat{\mathcal{D}}$ that are close to $\mathcal{D}$, the algorithm succeeds *with high probability* over a sample $\boldsymbol{S}$ drawn from $\widehat{\mathcal{D}}$. On the other hand, an algorithm succeeds in the presence of adaptive

adversaries if *with high probability* over a sample $\boldsymbol{S}$ drawn from $\mathcal{D}$, the algorithm succeeds *for all* corruptions $\widehat{\boldsymbol{S}}$ that are close to $\boldsymbol{S}$. Our work formalizes the question of when these quantifiers can be swapped, and our results provide several answers.

Future work In this work we initiate the systematic study of the power of adaptivity in statistical adversaries. A concrete direction for future work is to answer Question 1 for other broad classes of algorithms and natural noise models, either via the subsampling filter (Conjecture 1) or otherwise. Here we highlight the specific case of subtractive noise: having resolved the case of additive noise in this work, doing so for subtractive noise as well would be a significant step towards resolving Question 1 for all three generic noise models described in Section 3.

2. Preliminaries

We use **boldface** (e.g. $\boldsymbol{x} \sim \mathcal{D}$) to denote random variables. Throughout this paper $\mathcal{X}$ denotes an arbitrary finite domain, and we write $S \in \mathcal{X}^*$ to represent a multiset of elements in $\mathcal{X}$, meaning $S \in \mathcal{X}^0 \cup \mathcal{X}^1 \cup \mathcal{X}^2 \ldots$. We use the notation $a = b \pm \varepsilon$ to indicate that $|a - b| < \varepsilon$. For any $m \in \mathbb{N}$, the notation $[m]$ indicates the set $\{1, 2, \ldots, m\}$.

Distributions. For any $S \in \mathcal{X}^*$, we use $\mathcal{U}(S)$ to refer to the uniform distribution over S. For simplicity, we enforce that all distributions only have rational probabilities, meaning $\Pr_{\boldsymbol{x}\sim\mathcal{D}}[\boldsymbol{x} = x]$ is rational for any distribution $\mathcal{D}$ and element $x \in \mathcal{X}$.[6] For any distributions $\mathcal{D}_1, \mathcal{D}_2$ and parameter $\theta \in [0, 1]$, we use $\theta\mathcal{D}_1 + (1-\theta)\mathcal{D}_2$ to refer to the mixture distribution which samples from $\mathcal{D}_1$ with probability θ and from $\mathcal{D}_2$ with probability $(1 - \theta)$

Definition 2 (Total variation distance) *Let $\mathcal{D}$ and $\mathcal{D}'$ be any two distributions over the same domain, $\mathcal{X}$. It is well known that the following are equivalent definitions for the* total variation distance *between $\mathcal{D}$ and $\mathcal{D}'$, denoted* $\mathrm{dist}_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}')$*:*

1. *It is characterized by the best test distinguishing the two distributions:*

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

2. *It is characterized by the coupling which makes the two random variables different with the smallest probability.*

$$\mathrm{dist}_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}') \coloneqq \inf_{(\boldsymbol{x},\boldsymbol{x}') \text{ a coupling of } \mathcal{D},\mathcal{D}'} \left\{ \Pr[\boldsymbol{x} \neq \boldsymbol{x}'] \right\}.$$

3. Adversarial noise models

To reason generally about various noise adversarial models, we represent the types of allowable corruptions by a budget η and cost function ρ, which maps each ordered pair of distributions to some non-negative cost (or infinity). The cost need not be symmetrical.

6. Alternatively, one could enforce that the cost function smoothly interpolate to irrational probabilities, which is the case for all standard noise models.

Definition 3 ((ρ, η)-oblivious adversary) *Given some cost function ρ and budget η, an algorithm operating in the oblivious adversary model will receive the following input: If the true data distribution is $\mathcal{D}$, then the algorithm will receive iid samples from an adversarial chosen $\widehat{\mathcal{D}}$ satisfying $\rho(\mathcal{D}, \widehat{\mathcal{D}}) \leq \eta$.*

Definition 4 ((ρ, η)-adaptive adversary) *Given some cost function ρ and budget η, an n-sample algorithm operating in adaptive adversary model will receive the following input: If the true data distribution is $\mathcal{D}$, first a clean sample $\boldsymbol{S} \sim \mathcal{D}^n$ is generated, and then the algorithm will get an adversarially chosen $\widehat{\boldsymbol{S}}$ satisfying $\rho(\mathcal{U}(\boldsymbol{S}), \mathcal{U}(\widehat{\boldsymbol{S}})) \leq \eta$.*

Throughout this paper, we use $\widehat{\square}$ to denote the corrupted version of a set or distribution.

We remark that the adaptive adversary as defined in Definition 4 is slightly stronger than the definitions usually considered, in the sense that there is usually a bound on the size of $\widehat{S}$ the adversary is allowed to produce. For example, in the nasty noise model (Definition 7), the adversary is only allowed to change points in the sample, and so $|\widehat{S}| = |S|$. All of our results apply regardless of the size of $\widehat{S}$.

3.1. Standard noise models from the literature

In this subsection, we present three generic adversary models and show how they are special cases of our framework with an appropriate choice of cost function. Other standard models, and how they fit within our framework, are given in Appendix A.

Definition 5 (Additive noise) *Given a size-n sample $S \in \mathcal{X}^n$ and a corruption budget η, the adaptive additive noise adversary is allowed to add $\lfloor n \cdot \eta/(1-\eta) \rfloor$ points to S arbitrarily.*

The additive noise model is captured by the cost function:

$$\mathrm{cost}_{\mathrm{add}}(\mathcal{D}, \widehat{\mathcal{D}}) \coloneqq \inf_{\eta \in \mathbb{R}_{\geq 0}} \left\{ \widehat{\mathcal{D}} = (1-\eta)\mathcal{D} + \eta\mathcal{E} \right\} \quad \text{for some distribution } \mathcal{E}.$$

The oblivious version of additive noise is the well-known Huber contamination model (Huber, 1964).

Definition 6 (Subtractive noise) *Given a size-n sample $S \in \mathcal{X}^n$ and a corruption budget η, the adaptive subtractive noise adversary is allowed to remove $\lfloor \eta n \rfloor$ points from S arbitrarily.*

The subtractive noise adversary is captured by the cost function:

$$\mathrm{cost}_{\mathrm{sub}}(\mathcal{D}, \widehat{\mathcal{D}}) \coloneqq \mathrm{cost}_{\mathrm{add}}(\widehat{\mathcal{D}}, \mathcal{D}).$$

Definition 7 (Nasty noise) *Given a size-n sample $S \in \mathcal{X}^n$ and a corruption budget η, the adaptive nasty noise adversary is allowed to change up to $\lfloor \eta n \rfloor$ of the points in S arbitrarily.*

This noise model is also known as *strong contamination*, and as *nasty sample noise* (or simply nasty noise) in the context of supervised learning (Bshouty et al., 2002). It is captured by the cost function $\rho = \mathrm{dist}_{\mathrm{TV}}$.

4. Overview of Theorem 1: The SQ framework neutralizes adaptive adversaries

In this section we formally state and prove Theorem 1—namely that the behaviour of a Statistical Query (SQ) algorithm in the presence of adaptive adversaries is equivalent to its behaviour in the presence of oblivious adversaries.

Basics of the SQ framework. Let $\mathcal{D}$ be a distribution over some domain $\mathcal{X}$. A statistical query is a pair (ϕ, τ) where $\phi : \mathcal{X} \to [-1, 1]$ is the query and $\tau > 0$ is a tolerance parameter. These queries can be answered by a statistical query oracle $\text{STAT}_{\mathcal{D}}$ which, given an SQ (ϕ, τ), returns a value v equal to $\mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\phi(\boldsymbol{x})]$ up to an additive error of τ i.e. $v \in \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}[\phi(\boldsymbol{x})] \pm \tau$.

Throughout this section, we will use some convenient shorthand. For any distribution $\mathcal{D}$ and multiset $S \in \mathcal{X}^*$, we write

$$\phi(\mathcal{D}) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathcal{D}}[\phi(\boldsymbol{x})] \qquad \text{and} \qquad \phi(S) \coloneqq \phi(\mathcal{U}(S)) = \frac{1}{|S|}\sum_{x\in S}\phi(x).$$

A k-query statistical query algorithm, $\mathcal{A}$, is an algorithm that makes a sequence of statistical queries to the oracle $\text{STAT}_{\mathcal{D}}$ one by one, using the result of the previous queries to decide which statistical query to make next.

Definition 8 (k-query SQ Algorithm) *A k-query SQ algorithm $\mathcal{A}$ is a sequence of k SQs*

$$(\phi^{(1)}, \tau^{(1)}), (\phi^{(2)}_{v_1}, \tau^{(2)}), (\phi^{(3)}_{v_1,v_2}, \tau^{(3)}), \ldots, (\phi^{(k)}_{v_1,\ldots,v_{k-1}}, \tau^{(k)})$$

to $\text{STAT}_{\mathcal{D}}$, where $\mathcal{A}$'s choice of the $(i+1)$-st SQ can depend on $v_1, \ldots, v_i$ which are the answers of $\text{STAT}_{\mathcal{D}}$ to the previous i SQs. For notational simplicity, we make the standard assumption that all the τ's are the same.

Using mechanisms to implement $\text{STAT}_{\mathcal{D}}$. The SQ framework is a stylized model that cleanly facilitates theoretical analyses; It allows the algorithm designer to abstract away an algorithm's interaction with a random sample and instead *assume* $\phi(\mathcal{D})$ can be accessed up to $\pm\tau$ accuracy.

For an SQ algorithm to be useful, the $\text{STAT}_{\mathcal{D}}$ oracle must be implemented. This is done by a *mechanism* which uses a random sample $\boldsymbol{S} \sim \mathcal{D}^n$ to simulate $\text{STAT}_{\mathcal{D}}$ with high probability. The interaction between a mechanism and SQ algorithm is depicted in Figure 1.

1. Fix some k-query SQ algorithm $\mathcal{A}$ that is unknown to the mechanism $\mathcal{M}$ and a distribution $\mathcal{D}$ that is unknown to both $\mathcal{A}$ and $\mathcal{M}$.
2. Draw a sample $\boldsymbol{S} \sim \mathcal{D}^n$ that is revealed to $\mathcal{M}$ but not $\mathcal{A}$.
3. For $i = 1, \ldots, k$,
   (a) $\mathcal{A}$ chooses a query $\phi^{(i)}$ (as a function of responses to previous queries).
   (b) $\mathcal{M}$ chooses a response v_i for the query which is revealed to $\mathcal{A}$.

Figure 1: The interaction between a mechanism $\mathcal{M}$ and SQ algorithm $\mathcal{A}$.

Definition 9 ((τ, δ)-accurate mechanisms) *A mechanism $\mathcal{M}$ is (τ, δ)-accurate for k-query SQ algorithms, if for any distribution $\mathcal{D}$ and SQ algorithm $\mathcal{A}$, with probability at least $(1 - \delta)$ over the*

randomness of $\boldsymbol{S}$ and $\mathcal{M}$,

$$v_i = \phi^{(i)}(\mathcal{D}) \pm \tau \quad \textit{for all } i = 1, \ldots, k$$

where v_i and $\phi^{(i)}$ are defined as in Figure 1.

In this work, we focus on the τ-rounding mechanism.

Definition 10 (τ-rounding mechanism) *Given a sample $S \in \mathcal{X}^n$ and query ϕ, the τ-rounding mechanism, denoted $\mathcal{M}_\tau$, returns the answer $v = \text{round}(\phi(S), \tau)$ where $\text{round}(x, \tau)$ refers to x rounded to the nearest integer multiple of τ.*

Fact 1 (The τ-rounding mechanism is accurate) *For any $k \in \mathbb{N}$ and $\delta, \tau > 0$, the τ-rounding mechanism with a sample size of*

$$n = O\left(\frac{k\log(1/\tau) + \log(1/\delta)}{\tau^2}\right)$$

is (τ, δ) accurate for k-query SQ algorithms.

Proof For each query, the $\mathcal{M}_\tau$ can return one of only $O(1/\tau)$ possible values (after rounding). The i^{th} query is chosen as a function of $v_1, \ldots, v_{i-1}$, so there are at most $O(1/\tau)^{i-1}$ possible choices for the i^{th} query, and only $O(1/\tau)^k$ total unique queries $\mathcal{A}$ could choose. Using a Chernoff bound and union bound over all possible queries, the probability $\mathcal{A}$ asks a query, ϕ, where $|\phi(\boldsymbol{S}) - \phi(\mathcal{D})| \geq \tau/2$ is at most

$$\exp\left(-\Omega(\tau^2 n) + O(k\log(1/\tau))\right)$$

For the n given in Fact 1, that probability is at most δ. The desired result follows from triangle inequality and $|\text{round}(\phi(\boldsymbol{S}), \tau) - \phi(\boldsymbol{S})| \leq \tau/2$. ■

The existence of accurate mechanisms (as in Fact 1) is the key to the SQ framework: SQ algorithms can assume that they have access to a $\text{STAT}_{\mathcal{D}}$ oracle, because for modest sample sizes and tiny failure probabilities, mechanisms are a $\text{STAT}_{\mathcal{D}}$ oracle.

4.1. The SQ framework in the presence of adversarial noise.

The SQ framework naturally extends to oblivious adversaries.

Definition 11 (k-query SQ Algorithm with an oblivious adversary) *Fix a cost function ρ and budget η. An k-query SQ algorithm, $\mathcal{A}$, in the presence of a (ρ, η)-oblivious adversary is a sequence of k SQs*

$$(\phi^{(1)}, \tau), (\phi^{(2)}_{v_1}, \tau), (\phi^{(3)}_{v_1, v_2}, \tau), \ldots, (\phi^{(k)}_{v_1, \ldots, v_{k-1}}, \tau)$$

each of which are answered according to $\text{STAT}_{\widehat{\mathcal{D}}}$ for some $\widehat{\mathcal{D}}$ satisfying $\rho(\mathcal{D}, \widehat{\mathcal{D}}) \leq \eta$. This $\widehat{\mathcal{D}}$ is the same for all queries, but adversarially chosen. $\mathcal{A}$'s choice of the $(i+1)$-st SQ can depend on $v_1, \ldots, v_i$ which are the answers of $\text{STAT}_{\widehat{\mathcal{D}}}$ to the previous i SQs.

Once again, to run an SQ algorithm, the $\text{STAT}_{\widehat{\mathcal{D}}}$ oracle is implemented by a mechanism. Given a sample $\boldsymbol{S} \sim \widehat{\mathcal{D}}$, any mechanism satisfying Definition 9 will be able to simulate the $\text{STAT}_{\widehat{\mathcal{D}}}$ oracle.

In the presence of (ρ, η)-*adaptive* adversaries, first a clean sample $\boldsymbol{S} \sim \mathcal{D}^n$ is drawn, and then the adversary chooses an η-corruption $\widehat{\boldsymbol{S}}$ that is passed to the mechanism, as shown in Figure 2.

1. Fix some k-query SQ algorithm $\mathcal{A}$ that is unknown to the mechanism $\mathcal{M}$ and a distribution $\mathcal{D}$ that is unknown to both $\mathcal{A}$ and $\mathcal{M}$.
2. Draw a sample $\boldsymbol{S} \sim \mathcal{D}^n$ that is revealed to neither $\mathcal{A}$ or $\mathcal{M}$.
3. An adversary chooses an $\widehat{\boldsymbol{S}}$ that is η-close to $\boldsymbol{S}$ which is revealed to $\mathcal{M}$ but not $\mathcal{A}$.
4. For $i = 1, \ldots, k$,
   (a) $\mathcal{A}$ chooses a query $\phi^{(i)}$ (as a function of responses to previous queries).
   (b) $\mathcal{M}$ chooses a response v_i for the query which is revealed to $\mathcal{A}$.

Figure 2: The interaction between a mechanism $\mathcal{M}$ and SQ algorithm $\mathcal{A}$ in the presence of an adaptive adversary.

Our goal is to show that there are mechanisms that can simulate $\text{STAT}_{\widehat{\mathcal{D}}}$ for some $\widehat{\mathcal{D}}$ η-close to $\mathcal{D}$ given just the corrupted sample $\widehat{\boldsymbol{S}}$.

Definition 12 ((τ, δ)-accurate in the presence of adaptive noise) *Fix a cost function ρ, budget η. A mechanism $\mathcal{M}$ is said to be (τ, δ)-accurate for k-query SQ algorithms in the presence of adaptive noise, if for any distribution $\mathcal{D}$ and SQ algorithm $\mathcal{A}$, the following holds. With probability at least $1 - \delta$ over the randomness of $\boldsymbol{S} \sim \mathcal{D}^n$,*

$$v_i = \phi^{(i)}(\widehat{\mathcal{D}}) \pm \tau \quad \textit{for all } i = 1, \ldots, k$$

for some $\widehat{\mathcal{D}}$ η-close to $\mathcal{D}$, where v_i and $\phi^{(i)}$ are defined as in Figure 2. In particular, this holds regardless of how the adversary chooses $\widehat{\boldsymbol{S}}$.

If $\mathcal{A}$ succeeds given $\text{STAT}_{\widehat{\mathcal{D}}}$ for every distribution $\widehat{\mathcal{D}}$ that is η-close to $\mathcal{D}$ (i.e. $\mathcal{A}$ is resilient to oblivious adversaries), then, with high probability, $\mathcal{A}$ also succeeds in the presence of an adaptive adversary when using a mechanism satisfying Definition 12. We will show that the rounding mechanism meets Definition 12 whenever ρ is "reasonable" in the sense of the following definition; this property is easily satisfied by all standard cost functions.

Definition 13 (Closed under mixtures) *We say that ρ is* closed under mixtures *if for any distributions $\mathcal{D}_1, \mathcal{D}_2, \widehat{\mathcal{D}}_1, \widehat{\mathcal{D}}_2$ and $\theta \in (0, 1)$,*

$$\rho(\theta\mathcal{D}_1 + (1-\theta)\mathcal{D}_2, \theta\widehat{\mathcal{D}}_1 + (1-\theta)\widehat{\mathcal{D}}_2) \leq \max(\rho(\mathcal{D}_1, \widehat{\mathcal{D}}_1), \rho(\mathcal{D}_2, \widehat{\mathcal{D}}_2)).$$

Requiring that ρ is closed under mixtures enforces that the adaptive and oblivious adversaries "match up" in the sense of making the same types of changes. This is formalized in the following fact, for which we provide a short proof in Appendix F.

Fact 2 *Let ρ be closed under mixtures, $\mathcal{D}$ be a distribution over $\mathcal{X}$, η be a corruption budget, and $n \in \mathbb{N}$. Suppose that for all $S \in \mathcal{X}^n$, there is a corresponding $\widehat{S}$ satisfying $\rho(\mathcal{U}(S), \mathcal{U}(\widehat{S})) \leq \eta$. Let $\widehat{\mathcal{D}}$ be the distribution where $\boldsymbol{x} \sim \widehat{\mathcal{D}}$ is generated by: 1) Drawing $\boldsymbol{S} \sim \mathcal{D}^n$ and 2), drawing $\boldsymbol{x} \sim \mathcal{U}(\widehat{\boldsymbol{S}})$. Then, $\rho(\mathcal{D}, \widehat{\mathcal{D}}) \leq \eta$.*

Our quantitative bounds will depend on a parameter that is related to the types of corruption the adversaries can make. Suppose that the adaptive adversary is required to keep the size of the corrupted sample the same as the clean sample ($|\widehat{S}| = |S|$). In this case, we require that if two samples S_1 and S_2 differ in only one point, then for any $\widehat{S}_1$ that is η-close to S_1, there is some $\widehat{S}_2$ that is η-close to S_2 where $\widehat{S}_1$ and $\widehat{S}_2$ differ in only a small number of points. The following definition generalizes that notion to the case where the adversary can also change the number of points in the sample.

Definition 14 (ℓ-local) *For any $\ell > 0$, a cost function ρ with budget η is ℓ-local if for any distributions $\mathcal{D}_1, \mathcal{D}_2$ and η-corruption $\widehat{\mathcal{D}}_1$ of $\mathcal{D}_1$, there is some η-corruption $\widehat{\mathcal{D}}_2$ of $\mathcal{D}_2$ satisfying* $\mathrm{dist}_{\mathrm{TV}}(\widehat{\mathcal{D}}_1, \widehat{\mathcal{D}}_2) \leq \ell \cdot \mathrm{dist}_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2)$

All of the adversary models in Section 3 are 1-local with the exception of the η-subtractive noise, which is $\frac{1}{1-\eta}$-local. We encourage the reader to think of ℓ as a constant.

We are now ready to state the formal version of Theorem 1, which generalizes Fact 1 to the setting of adversarial noise.

Theorem 5 (Formal version of Theorem 1) *For any ℓ-local cost function, adversary budget η, $\delta, \tau > 0$, and $k \in \mathbb{N}$, the τ-rounding mechanism with a sample size of*

$$n = O\left(\frac{\ell^2(k\log(1/\tau) + \log(1/\delta))}{\tau^2}\right)$$

is (τ, δ) accurate for k-query SQ algorithm in the presence of adaptive noise.

Proof sketch. Here, we prove Theorem 5 contingent on Lemmas 1 and 2, which we prove in Appendix B.

First, we will prove the special case where $\mathcal{A}$ makes only a single query.[7]

Lemma 1 (Theorem 5 in the case of a single SQ) *Let $\Psi : \mathcal{X} \to [-1, 1]$ be a statistical query, $T \in [-1, 1]$, and suppose:*

$$\Psi(\widehat{\mathcal{D}}) \leq T \qquad \text{for all } \widehat{\mathcal{D}} \text{ that are } \eta\text{-close to } \mathcal{D}.$$

Then, for any $\tau > 0$ and sample size $n \in \mathbb{N}$, the probability over $\boldsymbol{S} \sim \mathcal{D}^n$ that there is some $\widehat{\boldsymbol{S}}$ that is η-close to $\boldsymbol{S}$ satisfying

$$\Psi(\widehat{\boldsymbol{S}}) \geq T + \frac{\tau}{2}$$

is at most $\exp\left(-\frac{\tau^2 n}{8\ell^2}\right)$.

7. To prove Theorem 5 in the case of a single statistical query, we could apply Lemma 1 twice: Once to bound how large the adversary can make $\Psi(\widehat{\boldsymbol{S}})$ and once to bound how small it can make $\Psi(\widehat{\boldsymbol{S}})$. Instead, we will directly apply Lemma 1 to prove the multi-query case of Theorem 5.

In order to prove Theorem 5, we want to bound the probability that for a random $\boldsymbol{S} \sim \mathcal{D}^n$, there is some corruption $\widehat{\boldsymbol{S}}$ for which $\mathcal{M}_\tau$ is not τ-accurate. In detail, that means, for

$$\begin{aligned} v_1 &\coloneqq \mathrm{round}(\phi^{(1)}(\widehat{\boldsymbol{S}}), \tau) \\ v_{i+1} &\coloneqq \mathrm{round}(\phi^{(i+1)}_{v_1,\ldots,v_i}(\widehat{\boldsymbol{S}}), \tau) \quad \text{for } i \in \{0, 1, \ldots, k-1\}, \end{aligned} \tag{1}$$

and, for every $\widehat{\mathcal{D}}$ that is η-close to $\mathcal{D}$, there is some i for which $|v_i - \phi^{(i)}_{v_1,\ldots,v_{i-1}}(\widehat{\mathcal{D}})| > \tau$. Consider a single possible choice for $v_1, \ldots, v_k$ (which also fixes the k-statistical queries, $\phi^{(1)}, \ldots, \phi^{(k)}$). We use the separating hyperplane theorem to reduce to the case of a single statistical query.

Lemma 2 (**k-query SQ algorithm to a single SQ**) *Fix any k statistical queries $\phi^{(1)}, \ldots, \phi^{(k)} : \mathcal{X} \to [-1, 1]$ and k values $v_1, \ldots, v_k$. Suppose that there is no $\widehat{\mathcal{D}}$ that is η-close to $\mathcal{D}$ satisfying*

$$\phi^{(i)}(\widehat{\mathcal{D}}) \in v_i \pm \tau \qquad \textit{for every } i \in [k].$$

Then there exists a single statistical query $\Psi : \mathcal{X} \to [-1, 1]$ and threshold T with the following properties.

1. *$\Psi(\widehat{\mathcal{D}}) \le T$ for every $\widehat{\mathcal{D}}$ that is η-close to $\mathcal{D}$.*
2. *For any sample $\widehat{S}$ satisfying $\phi^{(i)}(\widehat{S}) \in v_i \pm \frac{\tau}{2}$ for each $i \in [k]$, it is also true that $\Psi(\widehat{S}) \ge T + \frac{\tau}{2}$.*

Applying Lemmas 1 and 2, for any fixed choice of $v_1, \ldots, v_k$, the probability there is some $\widehat{\boldsymbol{S}}$ η-close to a sample $\boldsymbol{S} \sim \mathcal{D}^n$ satisfying Equation (1), and for every $\widehat{\mathcal{D}}$ that is η-close to $\mathcal{D}$, there is some i for which $|v_i - \phi^{(i)}_{v_1,\ldots,v_{i-1}}(\widehat{\mathcal{D}})| > \tau$ is only $\exp\left(-\frac{\tau^2 n}{8\ell^2}\right)$. Since each $v_i \in [-1, 1]$ is an integer multiple of τ, there are at most $(\frac{2}{\tau} + 1)^k$ many choices for $(v_1, \ldots, v_k)$. A union bound over all these choices completes the proof of Theorem 5.

Acknowledgments

We thank Adam Klivans and Greg Valiant for helpful conversations. We are grateful to Greg for allowing us to include Theorem 4 which was proved jointly with him.

Guy and Li-Yang are supported by NSF CAREER Award 1942123. Jane is supported by NSF Award CCF-2006664. Ali is supported by a graduate fellowship award from Knight-Hennessy Scholars at Stanford University.

References

Noga Alon, Omri Ben-Eliezer, Yuval Dagan, Shay Moran, Moni Naor, and Eylon Yogev. Adversarial laws of large numbers and optimal regret in online classification. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 447–455, 2021.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. *SIAM Journal on Computing*, 2021.

Amos Beimel, Haim Kaplan, Yishay Mansour, Kobbi Nissim, Thatchaphol Saranurak, and Uri Stemmer. Dynamic algorithms against an adaptive adversary: Generic constructions and lower bounds. In *Proceedings of the 54rd Annual ACM SIGACT Symposium on Theory of Computing*, 2022.

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

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.

Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Abhradeep Guha Thakurta. A separation result between data-oblivious and data-aware poisoning attacks. *Advances in Neural Information Processing Systems*, 34, 2021.

Ilias Diakonikolas and Daniel M Kane. Recent advances in algorithmic high-dimensional robust statistics. *arXiv preprint arXiv:1911.05911*, 2019.

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.

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

Nika Haghtalab, Tim Roughgarden, and Abhishek Shetty. Smoothed analysis with adaptive adversaries. In *Proceedings of the 62nd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2021.

Frank R. Hampel. A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, 42(6):1887 – 1896, 1971. doi: 10.1214/aoms/1177693054. URL https://doi.org/10.1214/aoms/1177693054.

Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *Proceedings of the 55th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 454–463, 2014.

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

Peter Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1), 1964.

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

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

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

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.

Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, pages 1588–1628, 2015.

John W Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver*, volume 2, pages 523–531, 1975.

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

Banghua Zhu, Jiantao Jiao, and Jacob Steinhardt. Generalized resilience and robust statistics. *arXiv*, abs/1909.08755, 2019. URL http://arxiv.org/abs/1909.08755.

Appendix A. Other standard noise models

Here, we list other standard noise models and show they fall under our framework.

Definition 15 (Nasty classification noise (Bshouty et al., 2002)) *Let the domain be $\mathcal{X} = X \times Y$. Given a size-n sample $S \in \mathcal{X}^n$ and a corruption budget η, the adaptive nasty classification noise adversary is allowed to choose $\lfloor \eta n \rfloor$ points and for each one, change it from (x, y) to $(x, \widehat{y})$ for arbitrary $\widehat{y} \in Y$.*

This model is captured by the cost function:

$$\mathrm{cost}_{\mathrm{agn}}(\mathcal{D}, \widehat{\mathcal{D}}) = \begin{cases} \infty & \text{if } \mathcal{D} \text{ and } \mathcal{D}' \text{ do not have the same marginal distribution over } \mathcal{X} \\ \mathrm{dist}_{\mathrm{TV}}(\mathcal{D}, \widehat{\mathcal{D}}) & \text{otherwise} \end{cases}$$

The $(\mathrm{cost}_{\mathrm{agn}}, \eta)$-oblivious adversary corresponds exactly to the well-studied agnostic learning model (Haussler, 1992; Kearns et al., 1994). Hence, Theorem 1 implies that nasty classification noise and the agnostic learning model are identical for SQ algorithms.

The final noise model that we discuss is defined with respect to an adversary that has intermediate adaptive power:

Definition 16 (Malicious noise (Valiant, 1985)) *In the malicious noise model where the adversary has corruption budget η, a sample is generated point-by-point. For each point, independently with probability $1 - \eta$, that point is $\boldsymbol{x} \sim \mathcal{D}$. Otherwise, the adversary is allowed to make that point an arbitrary $x \in \mathcal{X}$ with knowledge of the previous points sampled but not the future points.*

On the relationship between malicious noise and additive noise. The malicious noise adversary does not have full adaptivity, as when they decide what point to add, they only have knowledge of previous points sampled and not future ones. We now show how to encode the fully adaptive version of malicious noise, in which the adversary knows all points in the sample when deciding corruptions, in our framework.

We first augment the domain to $\mathcal{X}' = \mathcal{X} \cup \{\varnothing\}$ where $\varnothing$ will be used to indicate the adversary can change this point arbitrarily. We then let $\mathcal{D}'$ be the distribution satisfying, for each $x \in \mathcal{X}$:

$$\mathcal{D}' = (1-\eta)\mathcal{D} + \eta\mathcal{D}_{\varnothing}$$

where $\mathcal{D}_{\varnothing}$ is the distribution that always outputs $\varnothing$. We define the cost function to be

$$\mathrm{cost}_{\mathrm{mal}}(\mathcal{D}', \widehat{\mathcal{D}}') = \begin{cases} \infty & \text{if } \Pr_{\boldsymbol{x}\sim\widehat{\mathcal{D}}'}[\boldsymbol{x} = \varnothing] > 0 \\ \infty & \text{if } \Pr_{\boldsymbol{x}\sim\widehat{\mathcal{D}}'}[\boldsymbol{x} = x] < \Pr_{\boldsymbol{x}\sim\widehat{\mathcal{D}}}[\boldsymbol{x} = x] \text{ for any } x \in \mathcal{X} \\ 0 & \text{otherwise.} \end{cases}$$

Note that in order for $\widehat{\mathcal{D}}'$ to be a valid corruption of $\mathcal{D}'$ (i.e. $\mathrm{cost}_{\mathrm{mal}}(\mathcal{D}', \widehat{\mathcal{D}}') \neq \infty$), all of the probability mass of $\widehat{\mathcal{D}}'$ must be over $\mathcal{X}$, with none on $\varnothing$. The above provides an encoding of an adaptive adversary that is at least as powerful as malicious noise. To further understand the corresponding oblivious adversary, we note that, after fixing η,

$$\mathrm{cost}_{\mathrm{mal}}(\mathcal{D}', \widehat{\mathcal{D}}') \neq \infty \quad \text{if and only if} \quad \widehat{\mathcal{D}}' = (1-\eta)\mathcal{D} + \eta\mathcal{E} \quad \text{for some distribution } \mathcal{E}.$$

Hence, the oblivious adversary corresponding to malicious noise is the same as the oblivious adversary for additive noise. Theorem 2 implies that the adaptive and oblivious versions of additive noise, as well as malicious noise, are all equivalent.

A.1. Technical remarks

Fixed budget vs. variable budget. Consider the nasty noise model (Definition 7), corresponding to $\rho = \mathrm{dist}_{\mathrm{TV}}$ in our framework. Given a size-n clean sample $\boldsymbol{S}$, the adaptive adversary can choose any η-fraction of the points to change arbitrarily to create the corrupted sample $\widehat{\boldsymbol{S}}$. Often an alternative definition is used where the adversary is allowed to arbitrarily change m points in $\boldsymbol{S}$, where m can vary based on the specific sample $\boldsymbol{S}$, as long as the marginal distribution of m over samples $\boldsymbol{S} \sim \mathcal{D}^n$ is $\mathrm{Bin}(n, \eta)$. This definition is used by (Diakonikolas et al., 2019; Zhu et al., 2019) to show that the adaptive adversary can simulate any oblivious adversary. Technically, for our definition of an adaptive adversary with fixed budget η, this fact is not strictly true.[8] However, all our results are readily extendable to these slightly stronger adaptive adversaries with random-budgets.

We define an adversary model which encompasses those adversaries with variable budgets. First, the adversary will generate a corrupted data set $\widehat{\boldsymbol{S}}$. Then, it is allowed to change some points in that set to create $\widehat{\boldsymbol{S}}'$ as long as, most of the time, $\widehat{\boldsymbol{S}}'$ and $\widehat{\boldsymbol{S}}$ are close.

Definition 17 (Strong (ρ, η)-adaptive adversary) *Given a cost function ρ and budget η, an n-sample algorithm operating in the* strong (ρ, η)*-adaptive adversary model will receive as input the sample $\widehat{\boldsymbol{S}}'$ where*

8. For example, suppose $\mathcal{X} = \{0, 1\}$, $\mathcal{D}$ is the identically 0 distribution, and $\eta = 0.1$. If a size-10 sample is taken, under our adaptive definition, $\widehat{S}$ will never have more than a single 1. However, the oblivious adversary can choose $\widehat{\mathcal{D}}$ that returns 1 with probability 0.1 and as a result is a non-zero chance that two 1's appear in a size-10 sample.

1. *If the true data distribution is $\mathcal{D}$, first a clean sample $\boldsymbol{S} \sim \mathcal{D}^n$ is generated.*
2. *The adversary chooses a $\widehat{\boldsymbol{S}}$ satisfying $\rho(\mathcal{U}(\boldsymbol{S}), \mathcal{U}(\widehat{\boldsymbol{S}})) \leq \eta$.*
3. *The adversary chooses some $\widehat{\boldsymbol{S}}'$ where, over the randomness of the original sample and the adversaries decisions, the following holds.*

$$\Pr_{\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{S}}'} \left[\mathrm{dist}_{\mathrm{TV}}(\mathcal{U}(\widehat{\boldsymbol{S}}), \mathcal{U}(\widehat{\boldsymbol{S}}')) \geq t \right] \leq \exp(-O(nt^2)) \qquad \text{for all } t \in (0,1). \tag{2}$$

All of our upper bounds on the strength of adaptive adversaries also apply to the strong adaptive adversary. See Remark 1 for changes in the proof needed for Theorem 1. For Theorems 2 and 3, we can use an even weaker restriction on the adversary and only require that

$$\mathop{\mathbb{E}}_{\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{S}}'} \left[\mathrm{dist}_{\mathrm{TV}}(\mathcal{U}(\widehat{\boldsymbol{S}}), \mathcal{U}(\widehat{\boldsymbol{S}}')) \right] = O(1/\sqrt{n}) \tag{3}$$

in place of Equation (2). See Remark 4 for how Equation (3) can be used to make Theorems 2 and 3 work with strong adaptive adversaries.

Finite vs infinite domains. For our analyses, we assume that the domain, $\mathcal{X}$, is finite. Our goal is to understand when an algorithm that succeeds in the presence of an oblivious adversary implies an algorithm that succeeds in the presence of an adaptive adversary. Any algorithm that succeeds in the presence of an oblivious adversary can only read finitely many bits of each data point, effectively discretizing the domain.

That said, Theorem 1 also applies to infinite domains. If the domain is infinite, a more general definition of closed under mixtures is required in place of the simpler Definition 13

Definition 18 (Closed under mixtures, infinite domain) *We say that ρ is* closed under mixtures *if for any distributions $\mathcal{D}, \mathcal{D}'$, and coupling of $\boldsymbol{x} \sim \mathcal{D}$, $\boldsymbol{x}' \sim \mathcal{D}'$ and a latent variable $\boldsymbol{z}$ (over any domain),*

$$\rho(\mathcal{D}, \mathcal{D}') \leq \sup_{\boldsymbol{z}} \left\{ \rho((\boldsymbol{x} \mid \boldsymbol{z}), (\boldsymbol{x}' \mid \boldsymbol{z})) \right\}.$$

When the domain is finite, Definitions 13 and 18 are equivalent. When it is infinite, Definition 18 is needed to prove Fact 2. The remainder of the proof of Theorem 1 is identical.

Appendix B. Missing Lemmas from the proof of Theorem 5

We will use a few standard technical tools:

Fact 3 (Separating hyperplane theorem) *Let $A, B \in \mathbb{R}^k$ be disjoint, nonempty, and convex. There exists a nonzero vector $w \in \mathbb{R}^k$ and $T \in \mathbb{R}$ such that $a \cdot w \leq T$ and $b \cdot w \geq T$ for all $a \in A$ and $b \in B$.*

Fact 4 (McDiarmid's inequality) *Suppose that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the c-bounded difference property: for any $(x_1, \ldots, x_n), (x_1', \ldots, x_n') \in \mathcal{X}^n$ that differ on only on a single coordinate, f satisfies*

$$\left| f(x_1, \ldots, x_n) - f(x_1', \ldots x_n') \right| \leq c.$$

Then, for any $\tau > 0$ and any distribution $\mathcal{D}$ over $\mathcal{X}$,

$$\Pr_{\boldsymbol{S}\sim\mathcal{D}^n}[f(\boldsymbol{S}) - \mu \geq \tau] \leq \exp\left(-\frac{2\tau^2}{c^2 n}\right) \qquad \text{where} \quad \mu \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}[f(\boldsymbol{S})].$$

We prove the following two Lemmas, restated for convenience.

Lemma 1 (Theorem 5 in the case of a single SQ) *Let $\Psi : \mathcal{X} \to [-1, 1]$ be a statistical query, $T \in [-1, 1]$, and suppose:*

$$\Psi(\widehat{\mathcal{D}}) \leq T \qquad \text{for all } \widehat{\mathcal{D}} \text{ that are } \eta\text{-close to } \mathcal{D}.$$

Then, for any $\tau > 0$ and sample size $n \in \mathbb{N}$, the probability over $\boldsymbol{S} \sim \mathcal{D}^n$ that there is some $\widehat{\boldsymbol{S}}$ that is η-close to $\boldsymbol{S}$ satisfying

$$\Psi(\widehat{\boldsymbol{S}}) \geq T + \frac{\tau}{2}$$

is at most $\exp\left(-\frac{\tau^2 n}{8\ell^2}\right)$.

Proof For any sample $S \in \mathcal{X}^n$, define

$$f(S) \coloneqq \sup_{\widehat{\mathcal{U}(S)} \text{ is } \eta\text{-close to } \mathcal{U}(S)} \left\{\Psi(\widehat{\mathcal{U}(S)})\right\}.$$

We will show that f satisfies the $c \coloneqq \frac{2\ell}{n}$-bounded difference property. Consider any samples $S, S' \in \mathcal{X}^n$ that differ in only a single point. We will show that for every point in the set

$$\{\Psi(\widehat{\mathcal{U}(S)}) \mid \widehat{\mathcal{U}(S)} \text{ is } \eta\text{-close to } \mathcal{U}(S)\},$$

there is some point in the set

$$\{\Psi(\widehat{\mathcal{U}(S')}) \mid \widehat{\mathcal{U}(S')} \text{ is } \eta\text{-close to } \mathcal{U}(S')\}$$

that differs from it by at most $\pm\frac{2\ell}{n}$, and vice versa. This implies that f satisfies the $c \coloneqq \frac{2\ell}{n}$-bounded difference property.

As S and S' only differ in a single piont,

$$\mathrm{dist}_{\mathrm{TV}}(\mathcal{U}(S), \mathcal{U}(S')) \leq \frac{1}{n}.$$

Furthermore by Definition 14, for any $\widehat{\mathcal{U}(S)}$ that is η-close to $\mathcal{U}(S)$, there is some $\widehat{\mathcal{U}(S')}$ that is η-close to $\mathcal{U}(S')$ satisfying

$$\mathrm{dist}_{\mathrm{TV}}(\widehat{\mathcal{U}(S)}, \widehat{\mathcal{U}(S')}) \leq \frac{\ell}{n}.$$

Using the Definition 2 and the fact that the range of Ψ is a length-2 interval, the above implies that

$$\left|\Psi(\widehat{\mathcal{U}(S)}) - \Psi(\widehat{\mathcal{U}(S')})\right| \leq \frac{2\ell}{n}$$

proving that f satisfies the $\frac{2\ell}{n}$-bounded difference property. By McDiarmid's inequality, $f(\boldsymbol{S})$ concentrates around its mean. Lastly, we show that

$$\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}[f(\boldsymbol{S})] \le T. \tag{4}$$

Suppose for the sake of contradiction that $\mathbb{E}[f(\boldsymbol{S})] > T + \varepsilon$ for some $\varepsilon > 0$. For each $S \in \mathcal{X}^n$, let $\widehat{\mathcal{U}(S)}$ be η-close to $\mathcal{U}(S)$ and satisfy $\Psi(\widehat{\mathcal{U}(S)}) \ge f(S) - \varepsilon$ (which exists by the definition of f). We can define $\widehat{\mathcal{D}}$ as the distribution where, to sample $\boldsymbol{x} \sim \widehat{\mathcal{D}}$, we

1. Draw an i.i.d. sample $\boldsymbol{S} \sim \mathcal{D}^n$.
2. Draw $\boldsymbol{x} \sim \widehat{\mathcal{U}(\boldsymbol{S})}$ uniformly.

By Fact 2, $\widehat{\mathcal{D}}$ is η-close to $\mathcal{D}$. Then,

$$T \ge \Psi(\widehat{\mathcal{D}}) = \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}[\Psi(\widehat{\mathcal{U}(\boldsymbol{S})})] \ge \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}[f(\boldsymbol{S}) - \varepsilon] = \mathbb{E}[f(\boldsymbol{S})] - \varepsilon > T.$$

This is a contradiction, so Equation (4) holds. Lemma 1 follows from McDiarmid's inequality applied to f. ■

Lemma 2 (k-query SQ algorithm to a single SQ) *Fix any k statistical queries $\phi^{(1)}, \dots, \phi^{(k)} : \mathcal{X} \to [-1, 1]$ and k values $v_1, \dots, v_k$. Suppose that there is no $\widehat{\mathcal{D}}$ that is η-close to $\mathcal{D}$ satisfying*

$$\phi^{(i)}(\widehat{\mathcal{D}}) \in v_i \pm \tau \qquad \textit{for every } i \in [k].$$

Then there exists a single statistical query $\Psi : \mathcal{X} \to [-1, 1]$ and threshold T with the following properties.

1. *$\Psi(\widehat{\mathcal{D}}) \le T$ for every $\widehat{\mathcal{D}}$ that is η-close to $\mathcal{D}$.*
2. *For any sample $\widehat{S}$ satisfying $\phi^{(i)}(\widehat{S}) \in v_i \pm \frac{\tau}{2}$ for each $i \in [k]$, it is also true that $\Psi(\widehat{S}) \ge T + \frac{\tau}{2}$.*

Proof We'll actually prove a slightly more general result. We'll show that for any any *distribution* $\mathcal{E}$ satisfying $\phi^{(i)}(\mathcal{E}) \in v_i \pm \frac{\tau}{2}$ for each $i \in [k]$, it is also true that $\Psi(\mathcal{E}) \ge T + \frac{\tau}{2}$. Lemma 2 follows by setting $\mathcal{E} = \mathcal{U}(\widehat{S})$.

We define $A \in \mathbb{R}^d$ to be

$$A \coloneqq \left\{ \left(\phi^{(1)}(\widehat{\mathcal{D}}), \dots, \phi^{(k)}(\widehat{\mathcal{D}}) \right) \mid \widehat{\mathcal{D}} \text{ is } \eta\text{-close to } \mathcal{D} \right\},$$

which is convex since the cost function is closed under mixtures (Definition 13). We define B to be

$$B \coloneqq \left\{ b \in \mathbb{R}^k \mid b_i \in v_i \pm \tau \text{ for all } i \in [k] \right\},$$

which is convex since it is the intersection of halfspaces. By the assumptions of Lemma 2, A and B are disjoint. Let $w \in \mathbb{R}^k$ and $T \in R$ be the vector and threshold respectively guaranteed to exist by the separating hyperplane theorem, normalized so that $\|w\|_1 = 1$. We define

$$\Psi(x) \coloneqq \sum_{i\in[k]} w_i \cdot \phi^{(i)}(x).$$

As $\|w\|_1 = 1$ and the range of each $\phi^{(i)}(x) \in [-1,1]$ for each $x \in \mathcal{X}$, it is also true that $\Psi(x) \in [-1,1]$. We show that Ψ, T meet the two criteria of Lemma 2. The first criteria holds by the separating hyperplane theorem. The second criteria is equivalent to showing that any $b \in B_{\text{inner}}$ satisfies $b \cdot w \geq T + \frac{\tau}{2}$ where

$$B_{\text{inner}} := \left\{ b \in \mathbb{R}^k \mid b_i \in v_i \pm \tfrac{\tau}{2} \text{ for all } i \in [k] \right\}.$$

There must be a minimal point for $b \cdot w$ at a "corner" of B. Let $b^\star$ be such a minimal point (i.e. $(b^\star)_i = v_i + c_i \cdot \tau$ for $c_i \in \{\pm 1\}$). For any $b \in B_{\text{inner}}$

$$\begin{aligned}
b \cdot w &= b^\star \cdot w + \sum_{i \in [k]} w_i (b_i - b_i^\star) \\
&= b^\star \cdot w + \sum_{i \in [k]} |w_i| \cdot |b_i - b_i^\star| && (b^\star \text{ minimal for } b \cdot w \text{ over } b \in B) \\
&\geq b^\star \cdot w + \sum_{i \in [k]} |w_i| \cdot \tfrac{\tau}{2} && (|b_i - b_i^\star| \geq \tau/2 \text{ for } b \in B_{\text{inner}}) \\
&\geq T + \tfrac{\tau}{2}. && (\text{Separating hyperplane theorem, } \|w\|_1 = 1)
\end{aligned}$$

This implies the second criteria of Lemma 2. ■

Remark 1 *Theorem 5 also holds with strong adaptive adversaries (Definition 17 in Appendix A.1) rather than just adaptive adversaries, with the appropriate changes in constants. The proof only differs in Lemma 1. We then wish to bound the probability that the adversary can make $\Psi(\widehat{\boldsymbol{S}}') \geq T + \tau/2$, where $\widehat{\boldsymbol{S}}$ and $\widehat{\boldsymbol{S}}'$ are as in Definition 17. For constant ℓ,*

$$\begin{aligned}
\Pr[\Psi(\widehat{\boldsymbol{S}}') \geq T + \tau/2] &\leq \Pr_{\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{S}}'} \left[|\Psi(\widehat{\boldsymbol{S}}) - \Psi(\widehat{\boldsymbol{S}}')| \geq \tau/4 \right] + \Pr_{\widehat{\boldsymbol{S}}}[\Psi(\widehat{\boldsymbol{S}}) \geq T + \tau/4] \\
&\leq \Pr_{\widehat{\boldsymbol{S}}, \widehat{\boldsymbol{S}}'} \left[\text{dist}_{\text{TV}}(\mathcal{U}(\widehat{\boldsymbol{S}}), \mathcal{U}(\widehat{\boldsymbol{S}}')) \geq \tau/8 \right] + \exp(-O(\tau^2 n)) && (\text{Lemma 1}) \\
&\leq \exp(-O(\tau^2 n)). && (\text{Definition 17})
\end{aligned}$$

Once Lemma 1 is modified to handle strong adaptive adversaries, the remainder of the proof of Theorem 5 applies unchanged.

Appendix C. Proof of Theorem 3: If adaptivity can be neutralized, subsampling does it

Towards tackling Question 1 for all algorithms, we define the subsampling filter, a natural "wrapper algorithm" that operates only on samples and can be applied to any existing algorithm:

Definition 19 (Subsampling Filter) *Define the subsampling filter $\Phi_{m \to n} : \mathcal{X}^m \to \mathcal{X}^n$ that, given a set $S \in \mathcal{X}^m$, subsamples n elements $\boldsymbol{S}' \sim \mathcal{U}(S)^n$ and returns them. We will also write $\Phi_{* \to n}$ when the size of S is variable.*

Intuitively, by requesting a large number of points m and randomly subsampling points for the original algorithm, the filter should able to neutralise some of the power of the adaptive adversary, since the adversary cannot know which subsample the algorithm will receive. In this section we will prove Theorem 3 which, informally speaking, states that *if* the noise model is such that adaptivity can be neutralised, then subsampling does it. In the next section, we carry out this proof strategy for the specific case of additive noise and establish Theorem 2.

C.1. Definitions and the formal statement of Theorem 3

We begin by formalizing what it means for the behavior of an algorithm $\mathcal{A}$ in the presence of an oblivious adversary to be equivalent to that of an algorithm $\mathcal{A}'$ in the presence of an adaptive adversary. Roughly speaking, that corresponds to the range of acceptance probabilities $\mathcal{A}$ can have with all possible oblivious adversaries being close to the range of acceptance probabilities that $\mathcal{A}'$ can have with all possible adaptive adversaries.

Definition 20 *Fix a cost function ρ, budget $\eta \geq 0$, and distribution $\mathcal{D}$ over $\mathcal{X}$. For an algorithm $\mathcal{A} : \mathcal{X}^* \to \{0,1\}$, we define:*

$$\text{Oblivious-Max}_{\rho,\eta}(\mathcal{A},\mathcal{D},n) \coloneqq \sup_{\rho(\mathcal{D},\widehat{\mathcal{D}})\leq\eta} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\widehat{\mathcal{D}}^n} [\mathcal{A}(\boldsymbol{S})] \right\},$$
$$\text{Oblivious-Min}_{\rho,\eta}(\mathcal{A},\mathcal{D},n) \coloneqq \inf_{\rho(\mathcal{D},\widehat{\mathcal{D}})\leq\eta} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\widehat{\mathcal{D}}^n} [\mathcal{A}(\boldsymbol{S})] \right\},$$

the maximum and minimum acceptance probabilities of $\mathcal{A}$ given an obliviously corrupted $\mathcal{D}$. We similarly define the adaptive versions:

$$\text{Adaptive-Max}_{\rho,\eta}(\mathcal{A},\mathcal{D},m) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^m} \left[\sup_{\rho(\mathcal{U}(\boldsymbol{S}),\mathcal{U}(\widehat{\boldsymbol{S}}))\leq\eta} \left\{ \mathcal{A}(\widehat{\boldsymbol{S}}) \right\} \right],$$
$$\text{Adaptive-Min}_{\rho,\eta}(\mathcal{A},\mathcal{D},m) \coloneqq \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^m} \left[\inf_{\rho(\mathcal{U}(\boldsymbol{S}),\mathcal{U}(\widehat{\boldsymbol{S}}))\leq\eta} \left\{ \mathcal{A}(\widehat{\boldsymbol{S}}) \right\} \right].$$

Definition 21 (ε-equivalent) *Fix a cost function ρ and a budget $\eta \geq 0$. Let $\mathcal{A}, \mathcal{A}' : \mathcal{X}^* \to \{0,1\}$ be two algorithms. We say that $\mathcal{A}$ in the presence of (ρ,η)-oblivious adversaries is (n,m,ε)-equivalent to $\mathcal{A}'$ in the presence of (ρ,η)-adaptive adversaries if the following holds for all distributions $\mathcal{D}$ over $\mathcal{X}$:*

$$\text{Adaptive-Max}_{\rho,\eta}(\mathcal{A}',\mathcal{D},m) = \text{Oblivious-Max}_{\rho,\eta}(\mathcal{A},\mathcal{D},n) \pm \varepsilon,$$

and likewise for Min *instead of* Max*. If the algorithms $\mathcal{A}$ or $\mathcal{A}'$ are randomized, then these expectations are also over the randomness of the algorithms.*

We now state the formal version of Theorem 3:

Theorem 6 (Formal version of Theorem 3: If it is possible, subsampling does it) *Fix a cost function ρ and budget $\eta \geq 0$. Suppose that $\mathcal{A}$ and $\mathcal{A}'$ are algorithms where $\mathcal{A}$ in the presence of (ρ,η)-oblivious adversaries is (n,m,ε)-equivalent to $\mathcal{A}'$ in the presence of (ρ,η)-adaptive adversaries.*

Consider the subsampling algorithm $\mathcal{A}_{\mathrm{sub}} \coloneqq \mathcal{A} \circ \Phi_{*\to n}$ *which, given a sample* $S \in \mathcal{X}^*$*, subsamples* n *elements* $\boldsymbol{S}' \sim \mathcal{U}(S)^n$ *and returns* $\mathcal{A}(\boldsymbol{S}')$*. For*

$$M \coloneqq O\left(\frac{m^2 \log(1/\varepsilon)^2}{\varepsilon^5}\right),$$

we have that $\mathcal{A}$ *in the presence of* (ρ, η)*-oblivious adversaries is* $(n, M, 9\varepsilon)$*-equivalent to* $\mathcal{A}_{\mathrm{sub}}$ *in the presence of* (ρ, η)*-adaptive adversaries.*

Remark 2 (Search to decision reduction) *In both Theorems 2 and 3, we focus on decision algorithms that output a single bit* $\{0, 1\}$*, rather than the more general setting of search algorithms that output an answer from some set* $\mathcal{Y}$*. This is without loss of generality. Given a search algorithm* $\mathcal{A} : \mathcal{X}^* \to \mathcal{Y}$*, and any set of "good outputs"* $Y \subseteq \mathcal{Y}$*, we could define an algorithm* $\mathcal{B} \coloneqq \mathbb{1}_Y \circ \mathcal{A}$ *where* $\mathcal{B}(S) = 1$ *iff* $\mathcal{A}(S) \in Y$*. Then, we can directly apply Theorems 2 and 3 to* $\mathcal{B}$*. Hence, Theorems 2 and 3 hold for search algorithms with the appropriate definition of "equivalence" for search algorithms: We say* $\mathcal{A} : \mathcal{X}^\star \to \mathcal{Y}$ *and* $\mathcal{A}' : \mathcal{X}^\star \to \mathcal{Y}$ *are* ε*-equivalent if for every* $Y \subseteq \mathcal{Y}$*,* $\mathbb{1}_Y \circ \mathcal{A}$ *and* $\mathbb{1}_Y \circ \mathcal{A}'$ *are* ε*-equivalent (according to Definition 21).*

C.2. Proof of Theorem 6

Our proof of Theorem 6 relies on the following simple lemma. Roughly speaking, it states that sampling with replacement and sampling without replacement are nearly indistinguishable when the population is a quadratic factor larger than the number of samples.

Lemma 3 *For any distribution* $\mathcal{D}$ *and integers* $m, M \in \mathbb{N}$*, let* $\Phi_{M\to m} \circ \mathcal{D}^M$ *be the distribution with the following generative process: first draw a size-*M *sample* $\boldsymbol{S} \sim \mathcal{D}^M$*, and then subsample, with replacement,* m *points from* $\boldsymbol{S}$*. Then*

$$\mathrm{dist}_{\mathrm{TV}}(\mathcal{D}^m, \Phi_{M\to m} \circ \mathcal{D}^M) \le \frac{\binom{m}{2}}{M}.$$

Proof We describe a coupling of $\boldsymbol{S} \sim \mathcal{D}^m$ and $\boldsymbol{S}' \sim \Phi_{M\to m} \circ \mathcal{D}^M$ such that $\Pr[\boldsymbol{S} \neq \boldsymbol{S}'] \le \binom{m}{2}/M$.

1. Initialize S and S' to be empty sets, and $y_1, \ldots, x_M$ to be unset variables.
2. Repeat m times:
   (a) Draw $\boldsymbol{i} \sim [M]$ uniformly.
   (b) If $y_{\boldsymbol{i}}$ is unset, draw $\boldsymbol{x} \sim \mathcal{D}$ and set $y_{\boldsymbol{i}} \leftarrow \boldsymbol{x}$. Then, add $\boldsymbol{x}$ to both $\boldsymbol{S}$ and $\boldsymbol{S}'$.
   (c) Otherwise, add $y_{\boldsymbol{i}}$ to $\boldsymbol{S}'$ and sample $\boldsymbol{x} \sim \mathcal{D}$ to add to $\boldsymbol{S}$.

It is straightforward to verify that the above generative process leads to the distribution of $\boldsymbol{S}$ and $\boldsymbol{S}'$ being that of $\mathcal{D}^m$ and $\Phi_{M\to m} \circ \mathcal{D}^M$ respectively. Furthermore, if $\boldsymbol{S} \neq \boldsymbol{S}'$, that means there is some index $i \in [M]$ that was sampled at least twice. If we fix $j_1 \neq j_2 \in [m]$ and some $i \in [M]$, the probability i is the index chosen at steps j_1 and j_2 is $1/M^2$. Union bounding over the M choices for i and $\binom{m}{2}$ for j_1, j_2 gives that

$$\Pr[\boldsymbol{S} \neq \boldsymbol{S}'] \le \frac{\binom{m}{2}}{M}.$$

The desired result follows from the definition of total variation cost, Definition 2. ■

Lemma 3 will be used in conjunction with the following fact:

Fact 5 *Suppose there exists a test which, given c samples from a distribution $\mathcal{E}$ that is either $\mathcal{D}_0$ or $\mathcal{D}_1$, returns 0 if $\mathcal{E} = \mathcal{D}_0$ with probability at least $\frac{3}{4}$, and returns 1 if $\mathcal{E} = \mathcal{D}_1$ with probability at least $\frac{3}{4}$. Then,*

$$\operatorname{dist}_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2) \geq \frac{1}{2c}.$$

Together, Lemma 3 and Fact 5 imply that for appropriately chosen m and M, there is no sample-efficient test distinguishing $\mathcal{D}^m$ from $\Phi_{M\to m} \circ \mathcal{D}_M$ with high probability. We will use this to prove Theorem 6 by contradiction. Using the assumption that $\mathcal{A}'$ is equivalent to $\mathcal{A}$, we will design a sample-efficient test that approximates $\text{Oblivious-Max}_{\rho,\eta}$ and $\text{Oblivious-Min}_{\rho,\eta}$ for $\mathcal{A}$ with respect to both $\mathcal{D}^m$ and $\Phi_{M\to m} \circ \mathcal{D}^M$. We then show that if $\mathcal{A}_{\text{sub}} \coloneqq \Phi_{*\to n} \circ \mathcal{A}$ is *not* equivalent to $\mathcal{A}$, then these values will distinguish the two distributions.

The following lemma carries out the first part of this plan:

Lemma 4 *Let $\mathcal{A}$ and $\mathcal{A}'$ be as in Theorem 6. There is an estimator $\text{Est-Max}_{\rho,\eta}$ that uses*

$$m' \coloneqq \frac{m \log(2/\varepsilon)}{2\varepsilon^2}$$

samples from a distribution $\mathcal{D}$ and returns an estimate of $\text{Oblivious-Max}_{\rho,\eta}(\mathcal{A}, \mathcal{D}, n)$ that is accurate to $\pm 2\varepsilon$ with probability at least $1-\varepsilon$, and likewise an estimator $\text{Est-Min}_{\rho,\eta}$ for $\text{Oblivious-Min}_{\rho,\eta}$. Formally, for all distributions $\mathcal{D}$ over $\mathcal{X}$,

$$\Pr_{\boldsymbol{S}\sim\mathcal{D}^{m'}}\left[\text{Est-Max}_{\rho,\eta}(\boldsymbol{S}) = \text{Oblivious-Max}_{\rho,\eta}(\mathcal{A}, \mathcal{D}, n) \pm 2\varepsilon\right] \geq 1 - \varepsilon,$$

and likewise for $\text{Est-Min}_{\rho,\eta}$ and $\text{Oblivious-Min}_{\rho,\eta}$.

Proof We will prove the lemma for $\text{Est-Max}_{\rho,\eta}$ and $\text{Oblivious-Max}_{\rho,\eta}$; the proof for Min instead of Max is identical. $\text{Est-Max}_{\rho,\eta}$ computes an estimate satisfying:

$$\Pr_{\boldsymbol{S}\sim\mathcal{D}^{m'}}\left[\text{Est-Max}_{\rho,\eta}(\boldsymbol{S}) = \text{Adaptive-Max}_{\rho,\eta}(\mathcal{A}', \mathcal{D}, m) \pm \varepsilon\right] \leq \varepsilon. \tag{5}$$

This is sufficient to guarantee that $\text{Est-Max}_{\rho,\eta}$'s estimate is within $\pm 2\varepsilon$ of $\text{Oblivious-Max}_{\rho,\eta}(\mathcal{A}, \mathcal{D}, n)$ by our assumption that $\mathcal{A}$ is (n, m, ε)-equivalent to $\mathcal{A}'$. To provide such an estimate, $\text{Est-Max}_{\rho,\eta}(\boldsymbol{S})$ draws $\log(2/\varepsilon)/(2\varepsilon^2)$ many size-m samples $\boldsymbol{S}' \sim \mathcal{D}^m$. For each, it computes:

$$\sup_{\rho(\widehat{\boldsymbol{S}'}, \boldsymbol{S}') \leq \eta} \left\{\mathcal{A}'(\widehat{\boldsymbol{S}'})\right\} \tag{6}$$

and returns the average of these supremums. By the Chernoff bound, this average satisfies Equation (5). ■

Remark 3 *We are only concerned with the sample efficiency of these estimators, not their time efficiency or even whether they are computable. Indeed, as stated, an algorithm computing the estimators would need to loop or infinitely many $\widehat{\boldsymbol{S}'} \in \mathcal{X}^*$ to compute Equation (6). For us they are just an analytical tool used to prove Theorem 6, the conclusion of which gives an algorithm $\mathcal{A}_{\text{sub}} \coloneqq \mathcal{A} \circ \Phi_{*\to n}$ that inherits the time efficiency of $\mathcal{A}$.*

The next lemma notes that the Adaptive-Max$_{\rho,\eta}$ of $\mathcal{A}_{\text{sub}} \coloneqq \mathcal{A} \circ \Phi_{*\to n}$ can be expressed in terms of the Oblivious-Max$_{\rho,\eta}$ of $\mathcal{A}$. Formally:

Lemma 5 Adaptive-Max$_{\rho,\eta}(\mathcal{A}_{\text{sub}}, \mathcal{D}, M) = \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M}\left[\text{Oblivious-Max}_{\rho,\eta}(\mathcal{A}, \mathcal{U}(\boldsymbol{S}), n)\right]$, *and likewise for* Min *instead of* Max.

Proof The lemma follows from this series of identities:

$$\begin{aligned}
\text{Adaptive-Max}_{\rho,\eta}(\mathcal{A}_{\text{sub}}, \mathcal{D}, M) &= \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M}\left[\sup_{\rho(\mathcal{U}(\boldsymbol{S}),\mathcal{U}(\widehat{\boldsymbol{S}}))\le\eta}\left\{\mathcal{A}_{\text{sub}}(\widehat{\boldsymbol{S}})\right\}\right] \\
&\qquad\qquad \text{(Definition of Adaptive-Max}_{\rho,\eta}\text{)} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M}\left[\sup_{\rho(\mathcal{U}(\boldsymbol{S}),\mathcal{U}(\widehat{\boldsymbol{S}}))\le\eta}\left\{(\mathcal{A}\circ\Phi_{*\to n})(\widehat{\boldsymbol{S}})\right\}\right] \\
&\qquad\qquad \text{(Definition of } \mathcal{A}_{\text{sub}}\text{)} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M}\left[\sup_{\rho(\mathcal{U}(\boldsymbol{S}),\mathcal{E})\le\eta}\left\{\mathop{\mathbb{E}}_{\boldsymbol{S}'\sim\mathcal{E}^n}[\mathcal{A}(\boldsymbol{S}')]\right\}\right] \\
&\qquad\qquad \text{(Definition of the subsampling filter } \Phi_{*\to n}\text{)} \\
&= \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M}\left[\text{Oblivious-Max}_{\rho,\eta}(\mathcal{A}, \mathcal{U}(\boldsymbol{S}), n)\right], \\
&\qquad\qquad \text{(Definition of Oblivious-Max}_{\rho,\eta}\text{)}
\end{aligned}$$

where the penultimate identity also uses our convention that distributions have rational weights, and therefore can expressed as the uniform distribution over a sufficiently large multiset of elements. ■

The following lemma completes the proof of Theorem 6.

Lemma 6 *Let m' be as in Lemma 4 and define $M \coloneqq 14(m')^2/\varepsilon$. Then*

$$\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M}\left[\text{Oblivious-Max}_{\rho,\eta}(\mathcal{A}, \mathcal{U}(\boldsymbol{S}), n)\right] = \text{Oblivious-Max}_{\rho,\eta}(\mathcal{A}, \mathcal{D}, n) \pm 9\varepsilon,$$

and likewise for Min *instead of* Max.

Proof Our proof proceed by contradiction: assuming that $\mathbb{E}_{\boldsymbol{S}\sim\mathcal{D}^M}\left[\text{Oblivious-Max}_{\rho,\eta}(\mathcal{A}, \mathcal{U}(\boldsymbol{S}), n)\right]$ is more than 9ε far from Oblivious-Max$_{\rho,\eta}(\mathcal{A}, \mathcal{D}, n)$, we will prove that

$$\text{dist}_{\text{TV}}(\mathcal{D}^{m'}, \Phi_{M\to m'} \circ \mathcal{D}^M) \ge \frac{\varepsilon}{14}. \tag{7}$$

From Lemma 3, we know that

$$\text{dist}_{\text{TV}}(\mathcal{D}^{m'}, \Phi_{M\to m'} \circ \mathcal{D}^M) < \frac{(m')^2}{M} = \frac{\varepsilon}{14},$$

which yields the desired contradiction. To establish Equation (7), we design an algorithm that given $\lceil\frac{6}{\varepsilon}\rceil \le \frac{7}{\varepsilon}$ samples from either $\mathcal{D}^{m'}$ from $\Phi_{M\to m'} \circ \mathcal{D}^M$ is able to distinguish them with probability $\frac{3}{4}$. Once we do, the desired result follows from Fact 5.

Let $\mu \coloneqq \text{Oblivious-Max}_{\rho,\eta}(\mathcal{A}, \mathcal{D}, n)$. First, we define

$$\delta \coloneqq \Pr_{\boldsymbol{S}\sim\mathcal{D}^M}\left[|\text{Oblivious-Max}_{\rho,\eta}(\mathcal{A},\mathcal{U}(\boldsymbol{S}),n) - \mu| > 4\varepsilon\right]$$

and we bound

$$\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M}\left[\text{Oblivious-Max}_{\rho,\eta}(\mathcal{A},\mathcal{U}(\boldsymbol{S}),n)\right] \le (1-\delta)\cdot(\mu+4\varepsilon) + \delta\cdot 1 \le \mu + 4\varepsilon + \delta.$$

Similarly,

$$\begin{aligned}\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M}\left[\text{Oblivious-Max}_{\rho,\eta}(\mathcal{A},\mathcal{U}(\boldsymbol{S}),n)\right] &\ge (1-\delta)\cdot(\mu-4\varepsilon) + \delta\cdot 0 \\ &\ge \mu - 4\varepsilon - \delta. \qquad (\mu - 4\varepsilon \le 1)\end{aligned}$$

Hence by our assumption that

$$\left|\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M}\left[\text{Oblivious-Max}_{\rho,\eta}(\mathcal{A},\mathcal{U}(\boldsymbol{S}),n)\right] - \text{Oblivious-Max}_{\rho,\eta}(\mathcal{A},\mathcal{D},n)\right| > 9\varepsilon,$$

we can conclude that $\delta > 5\varepsilon$.

Now let $\mathcal{E}$ be either $\mathcal{D}^{m'}$ or $\Phi_{M\to m'}\circ\mathcal{D}^M$. Our test to determine which $\mathcal{E}$ is will do the following: draw $\lceil\frac{6}{\varepsilon}\rceil$ samples from $\mathcal{E}$, $\boldsymbol{S}\sim\mathcal{E}$, and run $\text{Est-Max}_{\rho,\eta}(\boldsymbol{S})$ on each. If less than a 2ε fraction of the estimates returned by $\text{Est-Max}_{\rho,\eta}(\boldsymbol{S})$ differ from μ in more than $\pm 2\varepsilon$, return that $\mathcal{E} = \mathcal{D}^{m'}$. Otherwise, return that $\mathcal{E} = \Phi_{M\to m'}\circ\mathcal{D}^M$. We will prove this test succeeds with probability at least $1 - e^{-2} \ge \frac{3}{4}$. We consider the two possible cases:

1. Case 1: $\mathcal{E} = \mathcal{D}^{m'}$. In this case, given a sample from $\mathcal{E}$, $\text{Est-Max}_{\rho,\eta}(\boldsymbol{S})$ returns an estimate that is within $\pm 2\varepsilon$ of μ with probability at least $1-\varepsilon$. By the Chernoff bound, given $\frac{6}{\varepsilon}$ such samples, the probability more than 2ε fraction deviate from μ by more than $\pm\varepsilon$ is at most $\exp(-\frac{1}{3}\cdot\frac{6}{\varepsilon}\cdot\varepsilon) = e^{-2}$. Therefore, the test succeeds with probability at least $1-e^{-2}$.

2. Case 2: $\mathcal{E} = \Phi_{M\to m'}\circ\mathcal{D}^M$. We showed above that with probability at least 5ε over a sample $\boldsymbol{S}\sim\mathcal{D}^M$ we have $|\text{Oblivious-Max}_{\rho,\eta}(\mathcal{A},\mathcal{U}(\boldsymbol{S}),n) - \mu| > 4\varepsilon$. When that's the case, $\text{Est-Max}_{\rho,\eta}(\boldsymbol{S})$ returns an estimate that is further than $\pm 2\varepsilon$ from μ with probability at least $1-\varepsilon$. Therefore, on a single sample, the probability that the estimate of $\mathcal{A}$ deviates from μ by more than $\pm 2\varepsilon$ is at least $5\varepsilon(1-\varepsilon) \ge 4\varepsilon$. By the Chernoff bound, given $\frac{6}{\varepsilon}$ samples, the probability that at most 2ε fraction deviate μ by at most $\pm 2\varepsilon$ is at most $\exp(-\frac{1}{8}\cdot\frac{6}{\varepsilon}\cdot 4\varepsilon) = e^{-3}$. Therefore, the test succeeds with probability at least $1-e^{-2}$.

Hence, given $\lceil\frac{6}{\varepsilon}\rceil$ samples, it is possible to distinguish $\mathcal{D}^{m'}$ from $\Phi_{M\to m'}\circ\mathcal{D}^M$ with a success probability of at least $\frac{3}{4}$. Equation (7) follows from Fact 5, completing the proof by contradiction. ■

Remark 4 (Strong adaptive adversaries) *If $\mathcal{A}: \mathcal{X}^n \to \{0,1\}$ in the presence of oblivious adversaries is (n, M, ε)-equivalent to $\mathcal{A}_{\text{sub}} \coloneqq \mathcal{A}\circ\Phi_{*\to n}$ in the presence of adaptive adversaries, then it is also $(n, M, 2\varepsilon)$-equivalent to $\mathcal{A}_{\text{sub}}$ in the presence of* strong adaptive adversaries *(Definition 17 in Appendix A.1) as long as $M = \Omega(n^2/\varepsilon^2)$. This applies to both Theorem 6 in this section and Theorem 7 in the next.*

Let $\widehat{\boldsymbol{S}}$ and $\widehat{\boldsymbol{S}}'$ be defined as in Definition 17. Then, for any strong adaptive adversary supplying the sample $\widehat{\boldsymbol{S}}'$, there is some adaptive adversary supplying the sample $\widehat{\boldsymbol{S}}$, so we wish to compare $\mathbb{E}[\mathcal{A}_{\mathrm{sub}}(\widehat{\boldsymbol{S}}')]$ to $\mathbb{E}[\mathcal{A}_{\mathrm{sub}}(\widehat{\boldsymbol{S}})]$. Recall that $\mathcal{A}_{\mathrm{sub}}$ first subsamples to n points, so

$$\begin{aligned}\left|\mathbb{E}[\mathcal{A}_{\mathrm{sub}}(\widehat{\boldsymbol{S}}')] - \mathbb{E}[\mathcal{A}_{\mathrm{sub}}(\widehat{\boldsymbol{S}})]\right| &\le n \mathop{\mathbb{E}}_{\widehat{\boldsymbol{S}},\widehat{\boldsymbol{S}}'}[\mathrm{dist}_{\mathrm{TV}}(\mathcal{U}(\widehat{\boldsymbol{S}}),\mathcal{U}(\widehat{\boldsymbol{S}}'))] \\ &\le O(n/\sqrt{M}) = \varepsilon. \qquad \text{(by Equation (3))}\end{aligned}$$

Appendix D. Proof of Theorem 2: The subsampling filter neutralizes adaptive additive noise

In this section, we prove the following theorem:

Theorem 7 (Formal version of Theorem 2)
Fix a budget $\eta \ge 0$, distribution $\mathcal{D}$ over $\mathcal{X}$, and algorithm $\mathcal{A} : \mathcal{X}^n \to \{0,1\}$. Consider the subsampling algorithm $\mathcal{A}_{\mathrm{sub}} \coloneqq \mathcal{A} \circ \Phi_{*\to n}$ which, given a sample $S \in \mathcal{X}^*$, subsamples n elements $\boldsymbol{S}' \sim \mathcal{U}(S)^n$ and returns $\mathcal{A}(\boldsymbol{S}')$. For

$$M \coloneqq O\left(\frac{n^4 \log(|\mathcal{X}|)}{\varepsilon^2}\right),$$

we have that $\mathcal{A}$ in the presence of $(\mathrm{cost}_{\mathrm{add}}, \eta)$-oblivious adversaries is (n, M, ε)-equivalent to $\mathcal{A}_{\mathrm{sub}}$ in the presence of $(\mathrm{cost}_{\mathrm{add}}, \eta)$-adaptive adversaries.

Implicit in the proof of Theorem 6, we proved the following.

Lemma 7 *Fix a cost function ρ, budget $\eta \ge 0$, and $\varepsilon > 0$. Suppose that for an algorithm $\mathcal{A} : \mathcal{X}^n \to \{0,1\}$, there are estimators $\text{Est-Max}_{\rho,\eta}, \text{Est-Min}_{\rho,\eta} : \mathcal{X}^m \to \{0,1\}$ that use m samples from a distribution $\mathcal{D}$ and returns estimates satisfying*

$$\Pr_{\boldsymbol{S}\sim\mathcal{D}^{m'}}\left[\text{Est-Max}_{\rho,\eta}(\boldsymbol{S}) = \text{Oblivious-Max}_{\rho,\eta}(\mathcal{A},\mathcal{D},n) \pm 2\varepsilon\right] \ge 1-\varepsilon,$$

and likewise for Min *instead of* Max. *Then, for*

$$M = \frac{14m^2}{\varepsilon},$$

we have that $\mathcal{A}$ in the presence of (ρ,η)-oblivious adversaries is $(n, M, 9\varepsilon)$-equivalent to $\mathcal{A}_{\mathrm{sub}} \coloneqq \mathcal{A} \circ \Phi_{*\to n}$ in the presence of (ρ,η)-adaptive adversaries.

In order to prove Theorem 7, we'll construct the estimator $\text{Est-Max}_{\mathrm{cost}_{\mathrm{add}},\eta}$ (and likewise for Min). The goal is to estimate

$$\text{Oblivious-Max}_{\mathrm{cost}_{\mathrm{add}},\eta}(\mathcal{A},\mathcal{D},n) \coloneqq \sup_{\widehat{\mathcal{D}}=(1-\eta)\cdot\mathcal{D}+\eta\cdot\mathcal{E}}\left\{\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\widehat{\mathcal{D}}^n}[\mathcal{A}(\boldsymbol{S})]\right\}.$$

The key insight is rather than trying all possible distributions $\mathcal{E}$, to compute a $\pm\varepsilon$ approximation, it suffices to consider those $\mathcal{E}$ that are equal to $\mathcal{U}(T)$ for some $T \in \mathcal{X}^{n^2/\varepsilon}$. Our final result will have a logarithmic dependence on the number of $\mathcal{E}$ we need to try, which results in just a logarithmic dependence on $|\mathcal{X}|$.

The following definition and accompany fact will be useful.

Definition 22 (Stochastic function) *For any sets $\mathcal{X}, \mathcal{Y}$, a stochastic function $\boldsymbol{f} : \mathcal{X} \to \mathcal{Y}$ is a collection of distributions $\{\mathcal{D}_x \mid x \in \mathcal{X}\}$ each supported on $\mathcal{Y}$ where the notation $\boldsymbol{f}(x)$ indicates an independent draw from $\mathcal{D}_x$.*

Fact 6 *For any distribution $\mathcal{D}_1, \mathcal{D}_2$ supported on a domain $\mathcal{X}$, and any stochastic function $\boldsymbol{f} : \mathcal{X} \to \mathcal{Y}$, let $\boldsymbol{f} \circ \mathcal{D}_i$ be the distribution where to sample $\boldsymbol{y} \sim \boldsymbol{f} \circ \mathcal{D}_i$, we first sample $\boldsymbol{x} \sim \mathcal{D}_i$ and then sample $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$. Then,*

$$\mathrm{dist}_{\mathrm{TV}}(\boldsymbol{f} \circ \mathcal{D}_1, \boldsymbol{f} \circ \mathcal{D}_2) \leq \mathrm{dist}_{\mathrm{TV}}(\mathcal{D}_1, \mathcal{D}_2).$$

Proof Given a coupling of $\boldsymbol{x}_1 \sim \mathcal{D}_1$ and $\boldsymbol{x}_2 \sim \mathcal{D}_2$, consider the coupling of $\boldsymbol{y}_1 \sim \boldsymbol{f} \circ \mathcal{D}_1$ and $\boldsymbol{y}_2 \sim \boldsymbol{f} \circ \mathcal{D}_2$ where if $\boldsymbol{x}_1 = \boldsymbol{x}_2$ then $\boldsymbol{y}_1 = \boldsymbol{y}_2 = \boldsymbol{f}(\boldsymbol{x}_1)$ and otherwise $\boldsymbol{y}_1 = \boldsymbol{f}(\boldsymbol{x}_1)$ and $\boldsymbol{y}_2 = \boldsymbol{f}(\boldsymbol{x}_2)$ independently. Then,

$$\Pr[\boldsymbol{y}_1 \neq \boldsymbol{y}_2] \leq \Pr[\boldsymbol{x}_1 \neq \boldsymbol{x}_2],$$

implying the desired result by Definition 2. ■

To prove Theorem 7, we will design the estimators $\text{Est-Max}_{\mathrm{cost}_{\mathrm{add}},\eta}$ and $\text{Est-Min}_{\mathrm{cost}_{\mathrm{add}},\eta}$. Plugging the sample complexity of those estimators into Lemma 7 would be sufficient for Theorem 7 with the slightly worse $M = n^6 \log(|\mathcal{X}|)^2/\varepsilon^7$. To get the optimal $\log(|\mathcal{X}|)$ dependence on the domain size (and an improved dependence on n and ε), we need a more refined version of Lemma 7 that takes advantage of the structure of the particular estimators we derive. The following Lemma applies for any cost function ρ, but we will then design $\mathcal{F}$ satisfying Equation (8) specifically for $\rho = \mathrm{cost}_{\mathrm{add}}$ in Lemma 9.

Lemma 8 *Fix a cost function ρ, budget $\eta \geq 0$, $\varepsilon \in (0, 1]$, and algorithm $\mathcal{A} : \mathcal{X}^n \to \{0, 1\}$. Suppose there is a set of stochastic functions $\mathcal{F}$ each $\mathcal{X} \to \mathcal{X}$, that satisfy, for any distribution $\mathcal{D}$ over $\mathcal{X}$,*

$$\max_{\boldsymbol{f} \in \mathcal{F}} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{S} \sim \mathcal{D}^n} [\mathcal{A}(\boldsymbol{f}(\boldsymbol{S}))] \right\} = \text{Oblivious-Max}_{\rho,\eta}(\mathcal{A}, \mathcal{D}, n) \pm \varepsilon, \tag{8}$$

and likewise for Min *instead of* Max*, where $\boldsymbol{f}(S)$ is shorthand for applying $\boldsymbol{f}$ element wise and independently to S. Then:*

1. *There are estimators $\text{Est-Max}_{\mathrm{cost}_{\mathrm{add}},\eta}$ and $\text{Est-Min}_{\mathrm{cost}_{\mathrm{add}},\eta}$ meeting the requirements of Lemma 7 for*

$$m' = O\left(\frac{n \log(|\mathcal{F}|/\varepsilon)}{\varepsilon^2}\right).$$

In particular, this implies that for

$$M = O\left(\frac{n^2 \log(|\mathcal{F}|/\varepsilon)^2}{\varepsilon^5}\right),$$

we have that $\mathcal{A}$ in the presence of (ρ, η)-oblivious adversaries is $(n, M, 9\varepsilon)$-equivalent to $\mathcal{A}_{\mathrm{sub}}$ in the presence of (ρ, η)-adaptive adversaries.

2. *More directly, for*

$$M = O\left(\frac{n^2 \log(|\mathcal{F}|/\varepsilon)}{\varepsilon}\right) \tag{9}$$

we have that $\mathcal{A}$ in the presence of (ρ, η)-oblivious adversaries is $(n, M, 5\varepsilon)$-equivalent to $\mathcal{A}_{\text{sub}}$ in the presence of (ρ, η)-adaptive adversaries.

Proof We first give $\text{Est-Max}_{\text{cost}_{\text{add}},\eta}$ satisfying the first item. For $r := O(\frac{\log(|\mathcal{F}|/\varepsilon)}{\varepsilon^2})$, let

$$\text{Est-Max}_{\text{cost}_{\text{add}},\eta} = \max_{\boldsymbol{f}\in\mathcal{F}} \left\{ \underbrace{\mathop{\mathbb{E}}_{\text{trial}\in[r]} \left[\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n} [\mathcal{A}(\boldsymbol{f}(\boldsymbol{S}))] \right]}_{:=\text{Est}^{(\boldsymbol{f})}} \right\}.$$

We note that as long as the samples are reused across different $\boldsymbol{f} \in \mathcal{F}$ that $m' = rn$ samples from $\mathcal{D}$ suffices to compute the above expression. By Hoeffding's inequality, for any fixed $\boldsymbol{f} \in \mathcal{F}$, with probability at least $1 - \frac{\varepsilon}{T}$

$$\text{Est}^{(\boldsymbol{f})} = \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n} [\mathcal{A}(\boldsymbol{f}(\boldsymbol{S}))] \pm \varepsilon.$$

By union bound, with probability at least $1-\varepsilon$, the above holds for every $\boldsymbol{f} \in \mathcal{F}$. If so, $\text{Est-Max}_{\text{cost}_{\text{add}},\eta}$ has the desired accuracy by Equation (8).

Next, we prove the M from Equation (9) suffices. By Lemma 5, it suffices to prove that

$$\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M} \left[\text{Oblivious-Max}_{\rho,\eta}(\mathcal{A}, \mathcal{U}(\boldsymbol{S}), n)\right] = \text{Oblivious-Max}_{\rho,\eta}(\mathcal{A}, \mathcal{D}, n) \pm 5\varepsilon,$$

and likewise for Min instead of Max. Applying Equation (8) to both sides of the above equation, it is sufficient to prove that

$$\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M} \left[\max_{\boldsymbol{f}\in\mathcal{F}} \left\{ \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{S}_n\sim\mathcal{U}(\boldsymbol{S})^n} [\mathcal{A}(\boldsymbol{f}(\boldsymbol{S}_n))]}_{:=g^{(\boldsymbol{f})}(\boldsymbol{S})} \right\} \right] = \max_{\boldsymbol{f}\in\mathcal{F}} \left\{ \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n} [\mathcal{A}(\boldsymbol{f}(\boldsymbol{S}))] \right\} \pm 3\varepsilon. \tag{10}$$

Fix a single $\boldsymbol{f} \in \mathcal{F}$. We'll use McDiarmid's inequality (Fact 4) to say that $g^{(\boldsymbol{f})}(\boldsymbol{S})$ concentrates around its mean. Take any $S \in \mathcal{X}^M$ and suppose we change one point in it to create S'. Then $|g^{(\boldsymbol{f})}(S) - g^{(\boldsymbol{f})}(S')|$ is at most the probability that the changed point appears in $\boldsymbol{S}_n$, which is at most $\frac{n}{M}$. Applying McDiarmid's inequality,

$$\begin{aligned} \Pr_{\boldsymbol{S}\sim\mathcal{D}^m} \left[g^{(\boldsymbol{f})}(\boldsymbol{S}) = \mu^{(\boldsymbol{f})} \pm \varepsilon \right] &\geq 1 - 2\exp\left(-\frac{2\varepsilon^2}{(n/M)^2 M}\right) && \text{where } \mu^{(\boldsymbol{f})} := \mathbb{E}_{\boldsymbol{S}\sim\mathcal{D}^M} \left[g^{(\boldsymbol{f})}(\boldsymbol{S}) \right] \\ &\geq 1 - \frac{\varepsilon}{|\mathcal{F}|}. && \left(\text{using } M = O\left(\tfrac{n^2 \log(|\mathcal{F}|/\varepsilon)}{\varepsilon}\right)\right) \end{aligned}$$

By union bound, with probability at least $1 - \varepsilon$, for every $\boldsymbol{f} \in \mathcal{F}$ we have that $g^{(\boldsymbol{f})}(\boldsymbol{S}) = \mu^{(\boldsymbol{f})} \pm \varepsilon$ allowing us to bound the left hand side of Equation (10),

$$\begin{aligned} \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^M} \left[\max_{\boldsymbol{f}\in\mathcal{F}} \left\{ g^{(\boldsymbol{f})}(\boldsymbol{S}) \right\} \right] &= \max_{\boldsymbol{f}\in\mathcal{F}} \left\{ \mu^{(\boldsymbol{f})} \right\} \pm \varepsilon \pm \Pr_{\boldsymbol{S}\sim\mathcal{D}^m} \left[g^{(\boldsymbol{f})}(\boldsymbol{S}) \neq \mu^{(\boldsymbol{f})} \pm \varepsilon \text{ for some } \boldsymbol{f} \in \mathcal{F} \right] \\ &= \max_{\boldsymbol{f}\in\mathcal{F}} \left\{ \mu^{(\boldsymbol{f})} \right\} \pm 2\varepsilon. \end{aligned}$$

Lastly, we want to compare the above to the right hand side of Equation (10) to $\max_{\boldsymbol{f}\in\mathcal{F}} \mu^{(\boldsymbol{f})}$. Fix any $\boldsymbol{f} \in \mathcal{F}$. Using the notation of Lemma 3,

$$\mu^{(\boldsymbol{f})} = \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\Phi_{M\to n}\circ\mathcal{D}^M}[\mathcal{A}(\boldsymbol{f}(\boldsymbol{S}))]\,.$$

Therefore,

$$\begin{aligned}\left|\max_{\boldsymbol{f}\in\mathcal{F}}\left\{\mu^{(\boldsymbol{f})}\right\} - \max_{\boldsymbol{f}\in\mathcal{F}}\left\{\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}[\mathcal{A}(\boldsymbol{f}(\boldsymbol{S}))]\right\}\right| &\le \max_{\boldsymbol{f}\in[\mathcal{F}]}\left|\mu^{(\boldsymbol{f})} - \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}[\mathcal{A}(\boldsymbol{f}(\boldsymbol{S}))]\right| \\ &\le \mathrm{dist}_{\mathrm{TV}}(\mathcal{D}^n, \Phi_{M\to n}\circ\mathcal{D}^M) && \text{(Definition 2)} \\ &\le \frac{\binom{n}{2}}{M} \le \varepsilon. && \text{(Lemma 3)}\end{aligned}$$

Therefore, the left hand term of Equation (10) is within $\pm 2\varepsilon$ of $\max_{\boldsymbol{f}\in\mathcal{F}} \mu^{(\boldsymbol{f})}$ and the right hand term is within $\pm\varepsilon$ of $\max_{\boldsymbol{f}\in\mathcal{F}} \mu^{(\boldsymbol{f})}$. Hence, Equation (10) holds, completing this proof. ■

In order to use Lemma 8, we need to design $\mathcal{F}$ and prove that it satisfies Equation (8) when $\rho = \mathrm{cost}_{\mathrm{add}}$. The below lemma completes the proof of Theorem 7.

Lemma 9 *Fix a budget $\eta \ge 0$, $\varepsilon \in (0, 1]$, and algorithm $\mathcal{A} : \mathcal{X}^n \to \{0, 1\}$. Let $\mathcal{F}$ be the set of $|\mathcal{X}|^{n^2/\varepsilon}$ stochastic functions defined below.*

$$\mathcal{F} \coloneqq \{\boldsymbol{f}^{(T)} \mid T \in \mathcal{X}^{n^2/\varepsilon}\} \qquad \textit{where} \quad \boldsymbol{f}^{(T)}(x) \coloneqq \begin{cases} x & \textit{with probability } 1-\eta \\ \boldsymbol{y} \textit{ where } \boldsymbol{y}\sim\mathcal{U}(T) & \textit{with probability } \eta. \end{cases}$$

Then for any $\mathcal{D}$ over $\mathcal{X}$,

$$\begin{aligned}\max_{\boldsymbol{f}\in\mathcal{F}}\left\{\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}[\mathcal{A}(\boldsymbol{f}(\boldsymbol{S}))]\right\} &= \text{Oblivious-Max}_{\mathrm{cost}_{\mathrm{add}},\eta}(\mathcal{A}, \mathcal{D}, n) \pm \varepsilon \\ &= \sup_{\widehat{\mathcal{D}}=(1-\eta)\cdot\mathcal{D}+\eta\cdot\mathcal{E}}\left\{\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\widehat{\mathcal{D}}^n}[\mathcal{A}(\boldsymbol{S})]\right\} \pm \varepsilon,\end{aligned}$$

and likewise for Min *instead of* Max.

Proof Fix any distribution $\mathcal{D}$. First, we note that the distribution of $\boldsymbol{f}^{(T)}(\boldsymbol{x})$ where $\boldsymbol{x} \sim \mathcal{D}$ is simply that of $(1-\eta)\cdot\mathcal{D} + \eta\cdot\mathcal{U}(T)$. Therefore, one direction of the desired result is easy:

$$\max_{\boldsymbol{f}\in\mathcal{F}}\left\{\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}[\mathcal{A}(\boldsymbol{f}(\boldsymbol{S}))]\right\} \le \text{Oblivious-Max}_{\mathrm{cost}_{\mathrm{add}},\eta}(\mathcal{A}, \mathcal{D}, n).$$

The remainder of this proof is devoted to proving the left hand side of the above equation is at most ε smaller than the right hand side. Fix any distribution $\mathcal{E}$ and consider $\widehat{\mathcal{D}} = (1-\eta)\cdot\mathcal{D} + \eta\cdot\mathcal{E}$. We'll show that

$$\mathop{\mathbb{E}}_{\boldsymbol{T}\sim\mathcal{E}^{n^2/\varepsilon}}\left[\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}[\mathcal{A}(\boldsymbol{f}^{\boldsymbol{T}}(\boldsymbol{S}))]\right] \ge \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\widehat{\mathcal{D}}^n}[\mathcal{A}(\boldsymbol{S})] - \varepsilon. \tag{11}$$

In particular, the above implies there is a single choice for T that is within ε of $\mathbb{E}_{\boldsymbol{S}\sim\widehat{\mathcal{D}}^n}[\mathcal{A}(\boldsymbol{S})]$. As this holds for all corruptions $\widehat{\mathcal{D}}$, it implies the desired result.

To sample from $\widehat{\mathcal{D}}^n$ where $\widehat{\mathcal{D}} = (1-\eta)\cdot\mathcal{D} + \eta\cdot\mathcal{E}$ we can first draw $\boldsymbol{S}\sim\mathcal{E}^n$ and then return $\boldsymbol{h}(\boldsymbol{S})$, applied element wise, where:

$$\boldsymbol{h}(x) \coloneqq \begin{cases} \boldsymbol{y} \text{ where } \boldsymbol{y}\sim\mathcal{D} & \text{with probability } 1-\eta \\ x & \text{with probability } \eta. \end{cases}$$

Therefore, the distribution of $\boldsymbol{f}^{\boldsymbol{T}}(\boldsymbol{S})$ on the left hand side of Equation (11) is $\boldsymbol{h}\circ\Phi_{n^2/\varepsilon\to n}\circ\mathcal{E}^{n^2/\varepsilon}$ (using the notation of Lemma 3 and Fact 6). Finally, we prove Equation (11) (recalling that $\widehat{\mathcal{D}} = (1-\eta)\cdot\mathcal{D}+\eta\cdot\mathcal{E}$)

$$\begin{aligned}
\left|\mathop{\mathbb{E}}_{\boldsymbol{T}\sim\mathcal{E}^{n^2/\varepsilon}}\left[\mathop{\mathbb{E}}_{\boldsymbol{S}\sim\mathcal{D}^n}[\mathcal{A}(\boldsymbol{f}^{\boldsymbol{T}}(\boldsymbol{S}))]\right] - \mathop{\mathbb{E}}_{\boldsymbol{S}\sim\widehat{\mathcal{D}}^n}[\mathcal{A}(\boldsymbol{S})]\right| &\le \mathrm{dist}_{\mathrm{TV}}(\boldsymbol{h}\circ\Phi_{n^2/\varepsilon\to n}\circ\mathcal{E}^{n^2/\varepsilon}, \boldsymbol{h}\circ\mathcal{E}^n) \\
&\qquad\qquad \text{(Definition 2)} \\
&\le \mathrm{dist}_{\mathrm{TV}}(\Phi_{n^2/\varepsilon\to n}\circ\mathcal{E}^{n^2/\varepsilon}, \mathcal{E}^n) \qquad \text{(Fact 6)} \\
&\le \varepsilon. \qquad\qquad \text{(Lemma 3)}
\end{aligned}$$

■

Appendix E. Proof of Theorem 4: Lower bounds against the subsampling filter

In this section, we show a lower bound on m needed for the subsampling filter to work. Our lower bound holds in the setting of additive noise and therefore also shows that the dependence on $|\mathcal{X}|$ in Theorem 7 is optimal.

Theorem 8 (Formal version of Theorem 4) *For any sample size n, domain $\mathcal{X}$ with $|\mathcal{X}| = 2^d$ for an integer d, adversary budget η, and $\varepsilon > 0$, there exists an algorithm $\mathcal{A}:\mathcal{X}^n\to\{0,1\}$ and corresponding subsampled algorithm $\mathcal{A}_{\mathrm{sub}} \coloneqq \mathcal{A}\circ\Phi_{*\to n}$ for which $\mathcal{A}$ in the presence of $(\mathrm{cost}_{\mathrm{add}},\eta)$-oblivious adversaries is* not *$(n,m,1-\varepsilon)$-equivalent to $\mathcal{A}_{\mathrm{sub}}$ in the presence of $(\mathrm{cost}_{\mathrm{add}},\eta)$-adaptive adversaries for any $m = O_\eta(n\log|\mathcal{X}|/\log^2 n)$.*

Proof overview Without loss of generality, we can consider the domain to be the Boolean hypercube, $\mathcal{X} = \{\pm1\}^d$. Otherwise, we could map the domain to the hypercube. For an appropriate threshold t, we'll define

$$\mathcal{A}(x_1,\dots,x_n) = \begin{cases} 1 & \text{if for every } x_i, \text{ there is an } x_j \text{ with } j\neq i \text{ s.t. } \langle x_i, x_j\rangle \ge t, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathcal{D}$ be uniform over $\mathcal{X}$. For any $\varepsilon>0$, n, and d, we'll show that there is a choice of t such that:

1. Lemma 10: $\text{Oblivious-Max}_{\mathrm{cost}_{\mathrm{add}},\eta}(\mathcal{A},\mathcal{D},n) \le \varepsilon/2$, meaning, for any $\widehat{\mathcal{D}} = (1-\eta)\mathcal{D}+\eta\mathcal{E}$, it is the case that $\mathbb{E}_{\boldsymbol{S}\sim\widehat{\mathcal{D}}^n}[\mathcal{A}(\boldsymbol{S})] \le \varepsilon/2$.

2. Lemma 11: $\text{Adaptive-Max}_{\text{cost}_{\text{add}},\eta}(\mathcal{A}, \mathcal{D}, m) \geq 1 - \frac{\varepsilon}{2}$ whenever $m = O_\eta(n \log |\mathcal{X}| / \log^2 n)$, meaning that for $\boldsymbol{S} \sim \mathcal{D}^m$, the adaptive adversary can choose $\lfloor m \cdot \eta/(1-\eta) \rfloor$ points $\boldsymbol{T}$ to add to the sample for which

$$\mathbb{E}[\mathcal{A}_{\text{sub}}(\widehat{\boldsymbol{S}})] \geq 1 - \tfrac{\varepsilon}{2} \quad \text{where } \widehat{\boldsymbol{S}} = \boldsymbol{S} \cup \boldsymbol{T}.$$

Together, these prove that $\mathcal{A}$ in the presence of $(\text{cost}_{\text{add}}, \eta)$-oblivious adversaries is not $(n, m, 1-\varepsilon)$-equivalent to $\mathcal{A}_{\text{sub}}$ in the presence of $(\text{cost}_{\text{add}}, \eta)$-adaptive adversaries.

Lemma 10 *For any distribution $\mathcal{E}$ and $\widehat{\mathcal{D}} = (1-\eta)\mathcal{D} + \eta\mathcal{E}$,*

$$\mathop{\mathbb{E}}_{\boldsymbol{S} \sim \widehat{\mathcal{D}}^n} [\mathcal{A}(\boldsymbol{S})] < \varepsilon/2.$$

Proof First, we note that for any $x' \in \mathcal{X}$ and a clean sample $\boldsymbol{x} \sim \mathcal{D}$, the probability that $\langle \boldsymbol{x}, x' \rangle \geq t$ is small: by Hoeffding's inequality, $\Pr_{\boldsymbol{x} \sim \mathcal{D}}[\langle \boldsymbol{x}, x' \rangle \geq t] \leq \exp(-t^2/2d)$. Combining this with a simple union bound, we can show that the probability of even a single clean point forming a correlated pair in the sample is small:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{S} \sim \widehat{\mathcal{D}}^n} [\mathcal{A}(\boldsymbol{S})] &\leq \eta^n + \mathop{\mathbb{E}}_{\boldsymbol{S} \sim \widehat{\mathcal{D}}^n} [\mathcal{A}(\boldsymbol{S}) \mid \text{at least one clean point in } \boldsymbol{S}] \\
&= \eta^n + \mathop{\mathbb{E}}_{\substack{\boldsymbol{x} \sim \mathcal{D} \\ \boldsymbol{S} \sim \widehat{\mathcal{D}}^{n-1}}} [\mathcal{A}(\boldsymbol{S} \cup \{\boldsymbol{x}\})] \\
&\leq \eta^n + \Pr_{\substack{\boldsymbol{x} \sim \mathcal{D} \\ \boldsymbol{S} \sim \widehat{\mathcal{D}}^{n-1}}} [\exists \boldsymbol{x}' \in \boldsymbol{S} \text{ with } \langle \boldsymbol{x}, \boldsymbol{x}' \rangle \geq t] && \text{(weaken to just one clean point)} \\
&\leq \eta^n + (n-1) \Pr_{\substack{\boldsymbol{x} \sim \mathcal{D} \\ \boldsymbol{x}' \sim \widehat{\mathcal{D}}}} [\langle \boldsymbol{x}, \boldsymbol{x}' \rangle \geq t] && \text{(union bound over } S\text{)} \\
&\leq \eta^n + n \exp\left(-\frac{t^2}{2d}\right). && \text{(Hoeffding's, over the randomness of } \boldsymbol{x}\text{)}
\end{aligned}$$

This will be vanishingly small for the particular choice of t determined in Lemma 11. ■

Lemma 11 *For any $m = O_\eta(nd/\log^2 n)$, there exists an adversarial strategy that, given $\boldsymbol{S} \in \mathcal{D}^m$, chooses $\lfloor m \cdot \eta/(1-\eta) \rfloor$ points $\boldsymbol{T}$ to add to the sample for which*

$$\mathbb{E}[\mathcal{A}_{\text{sub}}(\widehat{\boldsymbol{S}})] \geq 1 - \varepsilon/2 \quad \textit{where } \widehat{\boldsymbol{S}} = \boldsymbol{S} \cup \boldsymbol{T}$$

Proof Let $C = \lfloor m \cdot \eta/(1-\eta) \rfloor$ be the number of points the adversary can add and denote the sample $\boldsymbol{S} = \{\boldsymbol{x}^{(1)}, \ldots \boldsymbol{x}^{(m)}\}$. The adversary constructs $\boldsymbol{T} = \{\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(C)}\}$ by setting each $\boldsymbol{y}^{(j)}$ to be the elementwise majority of k chosen points from $\boldsymbol{S}$ (where k will be determined later). The idea is that $\boldsymbol{y}^{(j)}$ is a cluster center that will form a correlated pair with every one of these k points, with high probability.

More formally, for each $j = 1, \ldots, C$, define $\boldsymbol{S}^{(j)} = \{x^{(1+(j-1)k \pmod{|\boldsymbol{S}|})}, \ldots, x^{(jk \pmod{|\boldsymbol{S}|})}\}$ to be the j-th chunk of k points from $\boldsymbol{S}$, with the indices wrapping around to the start of $\boldsymbol{S}$ as necessary. We take $\boldsymbol{y}^{(j)}$ to be the elementwise majority[9] of the points in $\boldsymbol{S}^{(j)}$:

9. We can assume k is odd for simplicity.

$$\boldsymbol{y}_\ell^{(j)} \coloneqq \operatorname*{Maj}_{\boldsymbol{x} \in S^{(j)}} \{\boldsymbol{x}_\ell\} \qquad \text{for } \ell = 1, \dots, d. \tag{12}$$

First, we note that for a given $\boldsymbol{y}^{(j)}$, with high probability, any point in $\boldsymbol{x} \in \boldsymbol{S}^{(j)}$ will have a large dot product with $\boldsymbol{y}^{(j)}$:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{S}^{(j)} \sim \mathcal{D}^k}[\langle \boldsymbol{x}, \boldsymbol{y}^{(j)} \rangle] &= \sum_{\ell=1}^{d} \mathop{\mathbb{E}}_{\boldsymbol{S}^{(j)} \sim \mathcal{D}^k} \left[\boldsymbol{x}_\ell \boldsymbol{y}_\ell^{(j)} \right] \\
&= \sum_{\ell=1}^{d} \left(\Pr_{\boldsymbol{S}^{(j)} \sim \mathcal{D}^k} \left[\boldsymbol{x}_\ell = \boldsymbol{y}_\ell^{(j)} \right] - \Pr_{\boldsymbol{S}^{(j)} \sim \mathcal{D}^k} \left[\boldsymbol{x}_\ell \neq \boldsymbol{y}_\ell^{(j)} \right] \right) \\
&= \sum_{\ell=1}^{d} \left(\Pr_{\boldsymbol{u} \sim \mathrm{Bin}(k-1, \frac{1}{2})} \left[\boldsymbol{u} \geq \frac{(k-1)}{2} \right] - \Pr_{\boldsymbol{u} \sim \mathrm{Bin}(k-1, \frac{1}{2})} \left[\boldsymbol{u} < \frac{(k-1)}{2} \right] \right) \\
&= d \Pr_{\boldsymbol{u} \sim \mathrm{Bin}(k-1, \frac{1}{2})} \left[\boldsymbol{u} = \frac{(k-1)}{2} \right] \\
&= \sqrt{\frac{2}{\pi}} \frac{d}{\sqrt{k}} (1 \pm o(1)).
\end{aligned}$$

Define $\mu \coloneqq (\sqrt{2/\pi})(d/\sqrt{k})$. We will take $t = \mu/2$ as the threshold for $\mathcal{A}$, both here and in Lemma 10 as well. For $\boldsymbol{S}^{(j)} \sim \mathcal{D}^k$, with $\boldsymbol{y}^{(j)}$ the elementwise majority of $\boldsymbol{S}^{(j)}$, and any $\boldsymbol{x} \in S^{(j)}$, this gives:

$$\begin{aligned}
\cdot \Pr_{\boldsymbol{S}^{(j)} \sim \mathcal{D}^k} \left[\langle \boldsymbol{x}, \boldsymbol{y}^{(j)} \rangle < t \right] &= \Pr_{\boldsymbol{S}^{(j)} \sim \mathcal{D}^k} \left[\langle \boldsymbol{x}, \boldsymbol{y}^{(j)} \rangle < \frac{\mu}{2} \right] \\
&\leq \exp\left[-\Theta\left(\frac{\mu^2}{d} \right) \right] && \text{(Hoeffding's inequality)} \\
&= \exp\left[-\Theta\left(\frac{d}{k} \right) \right]. && (13)
\end{aligned}$$

The subsampling filter $\Phi_{* \to n}$ takes a random subsample of size n from $\widehat{\boldsymbol{S}} = \boldsymbol{S} \cup \boldsymbol{T}$ (with replacement). We want to show, with high probability over size n subsamples $\boldsymbol{S}' \sim \mathcal{U}(\widehat{\boldsymbol{S}})^n$, that $\mathcal{A}(\boldsymbol{S}') = 1$. For any point $\boldsymbol{x} \in \boldsymbol{S}$, we say $\boldsymbol{y}^{(j)} \in \boldsymbol{T}$ is "good" for $\boldsymbol{x}$ if $\boldsymbol{x}$ was in the cluster used to compute $\boldsymbol{y}^{(j)}$, meaning $\boldsymbol{x} \in \boldsymbol{S}^{(j)}$. Similarly, for any $\boldsymbol{y}^{(j)} \in \boldsymbol{T}$, we say that $\boldsymbol{x} \in \boldsymbol{S}$ is "good" for $\boldsymbol{y}^{(j)}$ is $\boldsymbol{x} \in \boldsymbol{S}^{(j)}$. By construction, a given $\boldsymbol{x} \in \boldsymbol{S}$ participates in the computation of at least $\lfloor Ck/m \rfloor = \Theta_\eta(k)$ many cluster centers $\boldsymbol{y}^{(j)}$'s, and so for each $\boldsymbol{x} \in \boldsymbol{S}$, there are $\Theta_\eta(k)$ good $\boldsymbol{y}$'s $\in T$. Similarly, there are exactly k good $\boldsymbol{x}$'s $\in S$ for each $\boldsymbol{y} \in T$. As $\widehat{\boldsymbol{S}}$ has size $\Theta_\eta(m)$, for any $\boldsymbol{x} \in \boldsymbol{S}'$, using $\boldsymbol{c}_{\boldsymbol{x}}$ to denote the number of good points for $\boldsymbol{x}$ that show up in $\boldsymbol{S}'$, we have that

$\boldsymbol{c_x}$ is distributed as $\mathrm{Bin}(n, \Omega_\eta(k/m))$. This gives:

$$\begin{aligned}
\Pr_{\boldsymbol{S}' \sim \mathcal{U}(\widehat{\boldsymbol{S}})^n}[\mathcal{A}(\boldsymbol{S}') = 0] &= \Pr_{\boldsymbol{S}' \sim \mathcal{U}(\widehat{\boldsymbol{S}})^n}[\exists \boldsymbol{x} \in \boldsymbol{S}' \text{ s.t. } \forall \boldsymbol{x}' \in \boldsymbol{S}', \langle \boldsymbol{x}, \boldsymbol{x}' \rangle < t] \\
&\le n \Pr_{\boldsymbol{x}, \boldsymbol{S}' \sim \mathcal{U}(\widehat{\boldsymbol{S}})^n}[\forall \boldsymbol{x}' \in \boldsymbol{S}', \langle \boldsymbol{x}, \boldsymbol{x}' \rangle < t] && \text{(union bound)} \\
&\le n \left[\Pr[\boldsymbol{c_x} = 0] + \Pr_{\boldsymbol{x}, \boldsymbol{S}' \sim \mathcal{U}(\widehat{\boldsymbol{S}})^n}[\forall \boldsymbol{x}' \in \boldsymbol{S}', \langle \boldsymbol{x}, \boldsymbol{x}' \rangle < t \mid \boldsymbol{c_x} \ge 1] \right] \\
&\le n \left[\Pr[\boldsymbol{c_x} = 0] + \Pr_{\boldsymbol{x}, \boldsymbol{S}' \sim \mathcal{U}(\widehat{\boldsymbol{S}})^n}[\langle \boldsymbol{x}, \boldsymbol{y_x} \rangle < t] \right] && \text{(weaken to good point)} \\
&\le n \left(1 - \Theta_\eta\left(\frac{k}{m}\right)\right)^n + n \exp\left[-\Theta_\eta\left(\frac{d}{k}\right)\right]. && \text{(from Equation (13))}
\end{aligned}$$

■

Setting of parameters. To complete the proof of Theorem 8, we need to choose m and k so that, all the terms in Lemma 10 and Lemma 11 are vanishingly small. In particular, we need $n\left(1 - \Theta_\eta\left(k/m\right)\right)^n \to 0$ and $n \exp\left(-\Theta_\eta\left(d/k\right)\right) \to 0$ as $n \to \infty$. If $k = \Omega_\eta(m \log n/n)$ and $d = \Omega_\eta(k \log n) = \Omega_\eta(m \log^2 n/n)$ with sufficiently large constant factors, we get the desired result. In other words, as long as $m \le O_\eta(nd/\log^2 n)$, the adaptive adversary is stronger than the oblivious adversary.

Appendix F. Proof of Fact 2

Proof By a standard inductive argument, Definition 13 implies that for any $m \in \mathbb{N}$, weights $\theta_1, \ldots, \theta_m \ge 0$ summing to 1 and distributions $\mathcal{D}_1, \ldots, \mathcal{D}_m, \widehat{\mathcal{D}}_1, \ldots, \widehat{\mathcal{D}}_m$, that

$$\rho\left(\sum_{i \in [m]} \theta_i \mathcal{D}_i, \sum_{i \in [m]} \theta_i \widehat{\mathcal{D}}_i\right) \le \max_{i \in [m]} \left\{\rho\left(\mathcal{D}_i, \widehat{\mathcal{D}}_i\right)\right\}.$$

The distribution $\mathcal{D}$ and $\widehat{\mathcal{D}}$ can be written as the mixtures

$$\begin{aligned}
\mathcal{D} &= \sum_{S \in \mathcal{X}^n} \Pr_{\boldsymbol{S} \sim \mathcal{D}^n}[\boldsymbol{S} = S] \mathcal{U}(S), \\
\widehat{\mathcal{D}} &= \sum_{S \in \mathcal{X}^n} \Pr_{\boldsymbol{S} \sim \mathcal{D}^n}[\boldsymbol{S} = S] \mathcal{U}(\widehat{S}).
\end{aligned}$$

Since $\rho(\mathcal{U}(S), \mathcal{U}(\widehat{S})) \le \eta$ for every $S \in \mathcal{X}^n$, we can conclude $\rho(\mathcal{D}, \widehat{\mathcal{D}}) \le \eta$. ■