Adaptive Data Analysis in a Balanced Adversarial Model

Kobbi Nissim* Uri Stemmer† Eliad Tsfadia‡

October 26, 2023

Abstract

In adaptive data analysis, a mechanism gets n i.i.d. samples from an unknown distribution $\mathcal{D}$, and is required to provide accurate estimations to a sequence of adaptively chosen statistical queries with respect to $\mathcal{D}$. Hardt and Ullman [HU14] and Steinke and Ullman [SU15b] showed that, in general, it is computationally hard to answer more than $\Theta(n^2)$ adaptive queries, assuming the existence of one-way functions.

However, these negative results strongly rely on an adversarial model that significantly advantages the adversarial analyst over the mechanism, as the analyst, who chooses the adaptive queries, also chooses the underlying distribution $\mathcal{D}$. This imbalance raises questions with respect to the applicability of the obtained hardness results – an analyst who has complete knowledge of the underlying distribution $\mathcal{D}$ would have little need, if at all, to issue statistical queries to a mechanism which only holds a finite number of samples from $\mathcal{D}$.

We consider more restricted adversaries, called *balanced*, where each such adversary consists of two separate algorithms: The *sampler* who is the entity that chooses the distribution and provides the samples to the mechanism, and the *analyst* who chooses the adaptive queries, but has no prior knowledge of the underlying distribution (and hence has no a priori advantage with respect to the mechanism).

We improve the quality of previous lower bounds by revisiting them using an efficient *balanced* adversary, under standard public-key cryptography assumptions. We show that these stronger hardness assumptions are unavoidable in the sense that any computationally bounded *balanced* adversary that has the structure of all known attacks, implies the existence of public-key cryptography.

*Department of Computer Science, Georgetown University. E-mail: kobbi.nissim@georgetown.edu. Work partially supported by NSF grant No. CNS-2001041 and a gift to Georgetown University.

†Blavatnik School of Computer Science, Tel Aviv University, and Google Research. E-mail: u@uri.co.il. Work partially supported by the Israel Science Foundation (grant 1871/19) and by Len Blavatnik and the Blavatnik Family foundation.

‡Department of Computer Science, Georgetown University. E-mail: eliadtsfadia@gmail.com. Work supported in part by the Fulbright Program and a gift to Georgetown University.

Contents

1 Introduction

Statistical validity is a widely recognized crucial feature of modern science. Lack of validity – popularly known as the *replication crisis* in science poses a serious threat to the scientific process and also to the public's trust in scientific findings.

One of the factors leading to the replication crisis is the inherent adaptivity in the data analysis process. To illustrate adaptivity and its effect, consider a data analyst who is testing a specific research hypothesis. The analyst gathers data, evaluates the hypothesis empirically, and often finds that their hypothesis is not supported by the data, leading to the formulation and testing of more hypotheses. If these hypotheses are tested and formed based on the same data (as acquiring fresh data is often expensive or even impossible), then the process is of *adaptive data analysis* (ADA) because the choice of hypotheses depends on the data. However, ADA no longer aligns with classical statistical theory, which assumes that hypotheses are selected independently of the data (and preferably before gathering data). ADA may lead to overfitting and hence false discoveries.

Statistical validity under ADA is a fundamental problem in statistics, that has received only partial answers. A recent line of work, initiated by [DFH+15c] and includes [HU14; DFH+15a; DFH+15b; SU15b; SU15a; BNS+16; RRST16; RZ16; Smi17; FS17; NSS+18; FS18; SL19; JLN+20; FRR20; DK22; KSS22; DSWZ23; Bla23] has resulted in new insights into ADA and robust paradigms for guaranteeing statistical validity in ADA. A major objective of this line of work is to design optimal mechanisms M that initially obtain a dataset $\mathcal{S}$ containing n i.i.d. samples from an unknown distribution $\mathcal{D}$, and then answers adaptively chosen queries with respect to $\mathcal{D}$. Importantly, all of M's answers must be accurate with respect to the underlying distribution $\mathcal{D}$, not just w.r.t. the empirical dataset $\mathcal{S}$. The main question is how to design an efficient mechanism that provides accurate estimations to adaptively chosen statistical queries, where the goal is to maximize the number of queries M can answer. This objective is achieved by providing both upper- and lower-bound constructions, where the lower-bound constructions demonstrate how an adversarial analyst making a small number of queries to an arbitrary M can invariably force M to err. The setting for these lower-bound proofs is formalized as a two-player game between a mechanism M and an adversary A as in Game 1.1.

Game 1.1 (ADA game between a mechanism M and an adversarial analyst A)**.**

- M *gets a dataset* $\mathcal{S}$ *of* n *i.i.d. samples from an* ***unknown*** *distribution* $\mathcal{D}$ *over* $\mathcal{X}$*.*
- *For* $i = 1, \ldots, \ell$*:*
  - A *sends a query* $q_i \colon \mathcal{X} \mapsto [-1, 1]$ *to* M*.*
  - M *sends an answer* $y_i \in [-1, 1]$ *to* A*.*
    (As A *and* M *are stateful,* q_i *and* y_i *may depend on the previous messages.)*

M *fails if* $\exists i \in [\ell]$ *s.t.* $|y_i - \mathrm{E}_{x \sim \mathcal{D}}[q_i(x)]| > 1/10$.

A question that immediately arises from the description of Game 1.1 is to whom should the distribution $\mathcal{D}$ be unknown, and how to formalize this lack of knowledge. Ideally, the mechanism M should succeed with high probability for every unknown distribution $\mathcal{D}$ and against any adversary A.

In prior work, this property was captured by letting the adversary choose the distribution $\mathcal{D}$ at the outset of Game 1.1. Namely, the adversary A can be seen as a pair of algorithms $(\mathsf{A}_1, \mathsf{A}_2)$, where A_1 chooses the distribution $\mathcal{D}$ and sends a state st to A_2 (which may contain the entire view of A_1), and after that, M and $\mathsf{A}_2(\mathsf{st})$ interacts in Game 1.1. In this adversarial model, Hardt and Ullman [HU14] and Steinke and Ullman [SU15b] showed that, assuming the existence of one way functions, it is computationally hard to answer more than $\Theta(n^2)$ adaptive queries. These results match the state-of-the-art constructions [DFH+15c; DFH+15a; DFH+15b; SU15a; BNS+16; FS17; FS18; DK22; Bla23].[1] In fact, each such negative result was obtained by constructing a *single* adversary A that fails *all* efficient mechanisms. This means that, in general, it is computationally hard to answer more than $\Theta(n^2)$ adaptive queries even when the analyst's algorithm is known to the mechanism. On the other hand, in each of these negative results, the adversarial analyst has a significant advantage over the mechanism – their ability to select the distribution $\mathcal{D}$. This allows the analyst to inject random trapdoors in $\mathcal{D}$ (e.g., keys of an encryption scheme) which are then used in forcing a computationally limited mechanism to fail, as the mechanism does not get a hold of the trapdoor information.

For most applications, the above adversarial model seems to be too strong. For instance, a data analyst who is testing research hypotheses usually has no knowledge advantage about the distribution that the mechanism does not have. In this typical setting, even if the underlying distribution $\mathcal{D}$ happens to have a trapdoor, if the analyst recovers the trapdoor then the mechanism should also be able to recover it and hence disable its adversarial usage.

In light of this observation, we could hope that in a balanced setting, where the underlying distribution is unknown to both the mechanism and the analyst, it would always be possible for M to answer more than $O(n^2)$ adaptive queries. To explore this possibility, we introduce what we call a *balanced* adversarial model.

Definition 1.2 (Balanced Adversary)**.** *A* balanced *adversary* A *consists of two isolated algorithms: The* sampler A_1*, which chooses a distribution* $\mathcal{D}$ *and provides i.i.d. samples to the mechanism* M*, and the* analyst A_2*, which asks the adaptive queries. No information is transferred from* A_1 *to* A_2*. See Game 1.3.*

Game 1.3 (The ADA game between a mechanism M and a balanced adversary $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)$)**.**

- A_1 *chooses a distribution* $\mathcal{D}$ *over* $\mathcal{X}$ *(specified by a sampling algorithm) and provides* n *i.i.d. samples* $\mathcal{S}$ *to* M *(by applying the sampling algorithm* n *times).*

  / A_1 *does not provide* A_2 *with any information */*

- M *and* A_2 *play Game 1.1 (with respect to* $\mathcal{D}$ *and* $\mathcal{S}$*).*

  M *fails if and only if it fails in Game 1.1.*

Note that the difference between the balanced model and the previous (imbalanced) one is

[1] Here is an example of a mechanism that handles $\tilde{\Theta}(n^2)$ adaptive queries using differential privacy: Given a query q_i, the mechanism returns an answer $y_i = \frac{1}{n}\sum_{x\in\mathcal{S}} x + \nu_i$ where the ν_i's are independent Gaussian noises, each with standard deviation of $\tilde{O}(\sqrt{\ell}/n)$. The noises guarantee that the entire process is "private enough" for avoiding overfitting in the ADA game, and accuracy is obtained whenever $\ell = \tilde{O}(n^2)$.

whether A_1 can send a state to A_2 after choosing the distribution $\mathcal{D}$ (in the imbalanced model it is allowed, in contrast to the balanced model).[2]

We remark that the main advantage of the balanced model comes when considering a publicly known sampler A_1 (as we do throughout this work). This way, A_1 captures the common knowledge that both the mechanism M and the analyst A_2 have about the underlying distribution $\mathcal{D}$.

Question 1.4. *Do the lower-bounds proved in prior work hold also for balanced adversaries?*

In this work we answer Question 1.4 in the positive. We do that using a publicly known analyst A_2 (which even makes it stronger than what is required for a lower bound). I.e., even though the sampler A_1 and the analyst A_2 are publicly known and cannot communicate with each other, they fail any computationally bounded mechanism. However, our lower-bound is based on stronger hardness assumptions than in prior work, namely, we use public-key cryptography.

1.1 Our Results

Our first result is a construction of a *balanced* adversary forcing any computationally bounded mechanim to fail in Game 1.3.

Theorem 1.5 (Computational lower bound, informal)**.** *There exists an efficient* balanced *adversary* $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)$ *that fails any computationally bounded mechanism* M *using* $\Theta(n^2)$ *adaptive queries. Moreover, it does so by choosing a distribution over a small domain.*

Our construction in Theorem 1.5 uses the structure of previous attacks of [HU14] and [SU15b], but relies on a stronger hardness assumption of public-key cryptography. We prove that this is unavoidable.

Theorem 1.6 (The necessity of public-key cryptography for proving Theorem 1.5, informal)**.** *Any computationally bounded* balanced *adversary that follows the structure of all currently known attacks, implies the existence of public-key cryptography (in particular, a* key-agreement *protocol).*

In Section 1.3 we provide proof sketches of Theorems 1.5 and 1.6, where the formal statements appear in Sections 3 and 4 (respectively) and the formal proofs appear in the supplementary material.

Potential Consequences for the Information Theoretic Setting. Theorem 1.6 has immediate implication to the information theoretic setting, and allow for some optimism regarding the possibility of constructing an inefficient mechanism that answers many adaptive queries.

It is known that an inefficient mechanism can answer exponentially many adaptive queries, but such results have a strong dependency on the domain size. For instance, the Private Multiplicative Weights algorithm of [HR10] can answer $2^{\tilde{O}\left(n/\sqrt{\log|\mathcal{X}|}\right)}$ adaptive queries accurately. However, this result is not useful whenever $n \leq O(\sqrt{\log|\mathcal{X}|})$. Indeed, [SU15b] showed that this dependency is unavoidable in general, by showing that large domain can be used for constructing a similar, unconditional, adversary that fails any computationally unbounded mechanism after $\Theta(n^2)$ queries.

[2] An additional (minor) difference is that we chose in our model to let A_1 also provide the i.i.d. samples to M. This is only useful for Theorem 1.6 as we need there that choosing $\mathcal{D}$ and sampling from $\mathcal{D}$ are both computationally efficient (which are simply captured by saying that A_1 is computationally bounded).

Our Theorem 1.6 implies that such an attack cannot be implemented in the balanced setting, which gives the first evidence that there might be a separation between the computational and information theoretic setting under the balanced adversarial model (in contrast with the imbalanced model).

Corollary 1.7. *There is no* balanced *adversary that follows the structure of all currently known attacks, and fails any (computationally* unbounded*) mechanism.*

In order to see why Corollary 1.7 holds, suppose that we could implement such kind of attack using a *balanced* adversary. Then by Theorem 1.6, this would imply that we could construct an information-theoretic key agreement protocol (i.e., a protocol between two parties that agree on a key that is secret from the eyes of a computationally *unbounded* adversary that only sees the transcript of the execution). But since the latter does not exist, we conclude that such a *balanced* adversary does not exists either. In other words, we do not have a negative result that rules out the possibility of constructing an inefficient mechanism that can answer many adaptive queries of a *balanced* adversary, and we know that if a negative result exists, then by Theorem 1.6 it cannot follow the structure of Hardt and Ullman [HU14] and Steinke and Ullman [SU15b].

1.2 Comparison with [Eld16]

The criticism about the lower bounds of Hardt and Ullman [HU14] and Steinke and Ullman [SU15b] is not new and prior work has attempted at addressing them with only partial success.

For example, Elder [Eld16] presented a similar "balanced" model (called "Bayesian ADA"), where both the analyst and the mechanism receive a *prior* $\mathcal{P}$ which is a family of distributions, and then the distribution $\mathcal{D}$ is drawn according to $\mathcal{P}$ (unknown to both the mechanism and the analyst).

From an information theoretic point of view, this model is equivalent to ours when the sampler A_1 is publicly known, since A_1 simply defines a prior. But from a computational point of view, defining the sampling process (i.e., sampling $\mathcal{D}$ and the i.i.d. samples from it) in an algorithmic way is better when we would like to focus on computationally bounded samplers.

[Eld16] only focused on the information-theoretic setting. His main result is that a certain family of mechanisms (ones that only use the posterior means) cannot answer more than $\tilde{O}(n^4)$ adaptive queries. This, however, does not hold for any mechanism's strategy. In particular, it does not apply to general computationally efficient mechanisms. Our negative result is quantitatively stronger (n^2 vs n^4) and it applies for all computationally efficient mechanisms.[3]

Table 1 summarizes the comparison between Theorem 1.5 and the prior lower bounds (ignoring computational hardness assumptions).

1.3 Techniques

We follow a similar technique to that used in [HU14] and [SU15b], i.e., a reduction to a restricted set of mechanisms, called *natural*.

Definition 1.8 (Natural mechanism [HU14])**.** *A mechanism* M *is* natural *if, when given a sample* $\mathcal{S} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ *and a query* $q\colon \mathcal{X} \to [-1, 1]$*,* M *returns an answer that is a function solely of* $(q(x_1), \ldots, q(x_n))$*. In particular,* M *does not evaluate* q *on other data points of its choice.*

[3] Our result is not directly comparable to that of [Eld16], because our negative result does not say anything for non-efficient mechanisms, while his result does rule out a certain family of non-efficient mechanisms.

	Balanced?	Class of Mechanisms	# of Queries	Dimension ($\log\|\mathcal{X}\|$)
[SU15b]	No	PPT Algorithms	$\tilde{O}(n^2)$	$n^{o(1)}$
[SU15b]	No	All	$\tilde{O}(n^2)$	$O(n^2)$
[Eld16]	Yes	Certain Family	$\tilde{O}(n^4)$	$\tilde{O}(n^4)$
Theorem 1.5	Yes	PPT Algorithms	$\tilde{O}(n^2)$	$n^{o(1)}$

Table 1: Comparison between the lower bounds for adaptive data analysis.

[HU14] and [SU15b] showed that there exists an adversarial analyst $\widetilde{\mathsf{A}}$ that fails any *natural* mechanism M, even when M is computationally unbounded, and even when $\mathcal{D}$ is chosen to be the uniform distribution over $\{1, 2, \ldots, m = 2000n\}$ (I.e., $\mathcal{D}$ is known to everyone). While general mechanisms could simply use the knowledge of the distribution to answer any query, *natural* mechanisms are more restricted, and can only provide answers based on the n-size dataset $\mathcal{S}$ that they get. The restriction to natural mechanisms allowed [SU15b] to use *interactive fingerprinting codes*, which enable to reveal $\mathcal{S}$ using $\Theta(n^2)$ adaptive queries when the answers are accurate and correlated with $\mathcal{S}$.

To construct an attacker A that fails any computationally bounded mechanism (and not just natural mechanisms), prior work forced the mechanism to behave naturally by using a private-key encryption scheme. More specifically, the adversary first samples m secret keys $\mathsf{sk}_1, \ldots, \mathsf{sk}_m$, and then defines $\mathcal{D}$ to be the uniform distribution over the pairs $\{(j, \mathsf{sk}_j)\}_{j=1}^m$. The adversary then simulates an adversary $\widetilde{\mathsf{A}}$ which fails natural mechanisms as follows: a query $\tilde{q}\colon [m] \to [-1, 1]$ issued by $\widetilde{\mathsf{A}}$ is translated by A to a set of m encryptions $\{\mathsf{ct}_j\}_{j=1}^m$ where each ct_j is an encryption of $\tilde{q}(j)$ under the key sk_j. These encryptions define a new query q that on input (j, sk), outputs the decryption of ct_j under the key sk. However, since M is computationally bounded and has only the secret keys that are part of its dataset $\mathcal{S}$, it can only decrypt the values of $\tilde{q}$ on points in $\mathcal{S}$, yielding that it effectively behaves *naturally*.

Note that the above attack A is *imbalanced* as it injects the secret keys $\mathsf{sk}_1, \ldots, \mathsf{sk}_m$ into $\mathcal{D}$ and then uses these keys when it forms queries. In other words, even though the attacker A is known to the mechanism, A is able to fail M by creating a secret correlation between its random coins and the distribution $\mathcal{D}$.

1.3.1 Balanced Adversary via Identity Based Encryption Scheme

For proving Theorem 1.5, we replace the private-key encryption scheme with a public-key primitive called *identity-based encryption* (IBE) scheme [Sha84; Coc01; BF01]. Such a scheme enables to produce m secret keys $\mathsf{sk}_1, \ldots, \mathsf{sk}_m$ along with a master public key mpk. Encrypting a message to a speific identity $j \in [m]$ only requires mpk, but decrypting a message for identity j must be done using its secret key sk_j. Using an IBE scheme we can achieve a reduction to *natural* mechanisms via a *balanced* adversary $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)$ as follows: A_1 samples keys $\mathsf{mpk}, \mathsf{sk}_1, \ldots, \mathsf{sk}_m$ according to the IBE scheme, and defines $\mathcal{D}$ to be the uniform distribution over the triplets $\{(j, \mathsf{mpk}, \mathsf{sk}_j)\}_{j=1}^m$. The analyst A_2, which does not know the keys, first asks queries of the form $q(j, \mathsf{mpk}, \mathsf{sk}) = \mathsf{mpk}_k$ for every bit k of mpk in order to reveal it. Then, it follows a strategy as in the previous section, i.e., it simulates an adversary $\widetilde{\mathsf{A}}$ which foils natural mechanisms by translating each query query $\tilde{q}\colon [m] \to [-1, 1]$ issued by $\widetilde{\mathsf{A}}$ by encrypting each $\tilde{q}(j)$ for identity j using mpk. Namely, the IBE scheme allowed the analyst to implement the attack of [HU14] and [SU15b], but without having to

know the secret keys $\mathsf{sk}_1, \ldots, \mathsf{sk}_m$.

We can implement the IBE scheme using a standard public-key encryption scheme: in the sampling process, we sample m independent pairs of public and secret keys $\{(\mathsf{pk}_j, \mathsf{sk}_j)\}_{j=1}^m$ of the original scheme, and define $\mathsf{mpk} = (\mathsf{pk}_1, \ldots, \mathsf{pk}_m)$. When encrypting a message for identity $j \in [m]$, we could simply encrypt it using pk_j (part of mpk), which can only be decrypted using sk_j. The disadvantage of this approach is the large master public key mpk that it induces. Applying the encryption scheme with security parameter of λ, the master key mpk will be of size $\lambda \cdot m$ and not just λ as the sizes of the secret keys. This means that implementing our *balanced* adversary with such an encryption scheme would result with a distribution over a large domain $\mathcal{X}$, which would not rule out the possibility to construct a mechanism for distributions over smaller domains. Yet, [DG21] showed that it is possible to construct a fully secure IBE scheme using a small mpk of size only $O(\lambda \cdot \log m)$ under standard hardness assumptions (e.g., the *Computational Diffie Helman* problem [DH76][4] or the hardness of *factoring*).

1.3.2 Key-Agreement Protocol via Balanced Adversary

In order to prove Theorem 1.6, we first explain what type of adversaries the theorem applies to. Recall that in all known attacks (including ours), the adversary A wraps a simpler adversary $\widetilde{\mathsf{A}}$ that fails *natural* mechanisms. In particular, the wrapper A has two key properties:

1. A knows $\mathrm{E}_{x\sim\mathcal{D}}[q_\ell(x)]$ for the last query q_ℓ that it asks (becuase it equals to $\frac{1}{m}\sum_{j=1}^m \tilde{q}_\ell(j)$, where $\tilde{q}_\ell$ is the wrapped query which is part of A's view), and

2. If the mechanism attempts to behave accurately in the first $\ell - 1$ rounds (e.g., it answers the empirical mean $\frac{1}{n}\sum_{x\in\mathcal{S}} q(x)$ for every query q), then A, as a wrapper of $\widetilde{\mathsf{A}}$, will be able to ask a last query q_ℓ that would fail any computationally bounded last-round strategy for the mechanism.

We next show that any computationally bounded *balanced* adversary A that has the above two properties, can be used for constructing a key-agreement protocol. That is, a protocol between two computationally bounded parties P_1 and P_2 that enable them to agree on a value which cannot be revealed by a computationally bounded adversary who only sees the transcript of the execution. See Protocol 1.9.

[4] CDH is hard with respect to a group $\mathbb{G}$ of order p, if given a random generator g along with g^a and g^b, for uniformly random $a, b \in [p]$, as inputs, the probability that a PPT algorithm can compute g^{ab} is negligible.

Protocol 1.9 (Key-Agreement Protocol $(\mathsf{P}_1, \mathsf{P}_2)$ via a *balanced* adversary $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)$)**.**

Input: 1^n*. Let* $\ell = \ell(n)$ *and* $\mathcal{X} = \mathcal{X}(n)$ *be the number queries and the domain that is used by the adversary* A*.*

Operation:

- P_1 *emulates* A_1 *on input* n *for obtaining a distribution* $\mathcal{D}$ *over* $\mathcal{X}$ *(specified by a sampling procedure), and samples* n *i.i.d. samples* $\mathcal{S}$*.*
- P_2 *initializes an emulation of* A_2 *on input* n*.*
- *For* $i = 1$ *to* ℓ*:*
  1. P_2 *receives the* i^{th} *query* q_i *from the emulated* A_2 *and sends it to* P_1*.*
  2. P_1 *computes* $y_i = \frac{1}{n}\sum_{x\in\mathcal{S}} q_i(x)$*, and sends it to* P_2*.*
  3. P_2 *sends* y_i *as the* i^{th} *answer to the emulated* A_2*.*
- P_1 *and* P_2 *agree on* $\mathrm{E}_{x\sim\mathcal{D}}[q_\ell(x)]$*.*

The agreement of Protocol 1.9 relies on the ability of P_1 and P_2 to compute $\mathrm{E}_{x\sim\mathcal{D}}[q_\ell(x)]$. Indeed, P_1 can accurately estimate it using the access to the sampling procedure, and P_2 can compute it based on the view of the analyst A_2 (follows by Property 1).

To prove the secrecy guarantee of Protocol 1.9, assume towards a contradiction that there exists a computationally bounded adversary G that given the transcript of the execution, can reveal $\mathrm{E}_{x\sim\mathcal{D}}[q_\ell(x)]$. Now consider the following mechanism for the ADA game: In the first $\ell - 1$ queries, answer the empirical mean, but in the last query, apply G on the transcript and answer its output. By the assumption on G, the mechanism will be able to accurately answer the last query, in contradiction to Property 2.

We note that Property 1 can be relaxed by only requiring that A is able to provide a "good enough" estimation of $\mathrm{E}_{x\sim\mathcal{D}}[q_\ell(x)]$. Namely, as long as the estimation provided in Property 1 is better than the estimation that an adversary can obtain in Property 2 (we prove that an $n^{\Omega(1)}$ multiplicative gap suffices), this would imply that Protocol 1.9 is a *weak* key-agreement protocol, which can be amplified to a fully secure one using standard techniques.

We also note that by requiring in Game 1.3 that A_1 samples from $\mathcal{D}$ according to the sampling procedure, we implicitly assume here that sampling from $\mathcal{D}$ can be done efficiently (because A_1 is assumed to be computationally bounded). Our reduction to key-agreement relies on this property, since if sampling from $\mathcal{D}$ could not be done efficiently, then P_1 would not have been a computationally bounded algorithm.

1.4 Perspective of Public Key Cryptography

Over the years, cryptographic research has proposed solutions to many different cryptographic tasks under a growing number of (unproven) computational hardness assumptions. To some extent, this state of affairs is unavoidable, since the security of almost any cryptographic primitive implies the existence of one-way functions [IL89] (which in particular implies that $P \neq NP$). Yet, all various assumptions can essentially be divided into two main types: *private key* cryptography and *public*

key cryptography [Imp95]. The former type is better understood: A series of works have shown that the unstructured form of hardness guaranteed by one-way functions is sufficient to construct many complex and structured primitives such as pseudorandom generators [HILL99], pseudorandom functions [GGM86] and permutations [LR88], commitment schemes [Nao91; HNO+09], universal one-way hash functions [Rom90], zero-knowledge proofs [GMW87], and more. However, reductions are much less common outside the one-way functions regime, particularly when constructing public-key primitives. In the famous work of Impagliazzo and Rudich [IR89] they gave the first evidence that *public key* cryptography assumptions are strictly stronger than one-way functions, by showing that key-agreement, which enables two parties to exchange secret messages over open channels, cannot be constructed from one-way functions in a black-box way.

Our work shows that a *balanced* adversary for the ADA game that has the structure of all known attacks, is a primitive that belongs to the public-key cryptography type. In particular, if public-key cryptography does not exist, it could be possible to construct a computationally bounded mechanism that can handle more than $\Theta(n^2)$ adaptive queries of a *balanced* adversary (i.e., we currently do not have a negative result that rules out this possibility).

1.5 Other Related Work

Nissim *et al.* [NSS+18] presented a variant of the lower bound of [SU15b] that aims to reduce the number of queries used by the attacker. However, their resulting lower bound only holds for a certain family of mechanisms, and it does not rule out all computationally efficient mechanisms.

Dinur *et al.* [DSWZ23] revisited and generalized the lower bounds of [HU14] and [SU15b] by showing that they are a consequence of a space bottleneck rather than a sampling bottleneck. Yet, as in the works by Hardt, Steinke, and Ullman, the attack by Dinur et al. relies on the ability to choose the underlying distribution $\mathcal{D}$ and inject secret trapdoors in it, and hence it utilizes an *imbalanced* adversary.

Recently, lower bounds constructions for the ADA problem were used as a tool for constructing (conditional) lower bounds for other problems, such as the space complexity of *adaptive streaming algorithms* [KMNS21] and the time complexity of *dynamic algorithms* [BKM+22]. Our lower bound for the ADA problem is qualitatively stronger than previous lower bounds (as the adversary we construct has less power). Thus, our lower bound could potentially yield new connections and constructions in additional settings.

1.6 Conclusion and Open Problems

In this work we present the balanced adversarial model for the ADA problem, and show that the existence of a balanced adversary that has the structure of all previously known attacks is equivalent to the existence of public-key cryptography. Yet, we do not know what is the truth outside of the public-key cryptography world. Can we present a different type of efficient attack that is based on weaker hardness assumptions (like one-way functions)? Or is it possible to construct an efficient mechanism that answer more than $\Theta(n^2)$ adaptive queries assuming that public-key cryptography does not exist? We also leave open similar questions regarding the information theoretic case. We currently do not know whether it is possible to construct an unbounded mechanism that answers exponential number of queries for any distribution $\mathcal{D}$ (regardless of its domain size).

In a broader perspective, lower bounds such as ours show that no general solution exists for a problem. They often use unnatural inputs or distributions and rely on cryptographic assumptions.

They are important as guidance for how to proceed with a problem, e.g., search for mechanisms that would succeed if the underlying distribution is from a "nice" family of distributions.

2 Preliminaries

2.1 Notations

We use calligraphic letters to denote sets and distributions, uppercase for random variables, and lowercase for values and functions. For $n \in \mathbb{N}$, let $[n] = \{1, 2, \ldots, n\}$. Let $\text{neg}(n)$ stand for a negligible function in n, i.e., a function $\nu(n)$ such that for every constant $c > 0$ and large enough n it holds that $\nu(n) < n^{-c}$. For $n \in \mathbb{N}$ we denote by 1^n the n-size string $1 \ldots 1$ (n times). Let PPT stand for probabilistic polynomial time. We say that a pair of algorithms $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)$ is PPT if both A_1 and A_2 are PPT algorithms.

2.2 Distributions and Random Variables

Given a distribution $\mathcal{D}$, we write $x \sim \mathcal{D}$, meaning that x is sampled according to $\mathcal{D}$. For a multiset $\mathcal{S}$, we denote by $\mathcal{U}_\mathcal{S}$ the uniform distribution over $\mathcal{S}$, and let $x \leftarrow \mathcal{S}$ denote that $x \sim \mathcal{U}_\mathcal{S}$. For a distribution $\mathcal{D}$ and a value $n \in \mathbb{N}$, we denote by $\mathcal{D}^n$ the distribution of n i.i.d. samples from $\mathcal{D}$. For a distribution $\mathcal{D}$ over $\mathcal{X}$ and a query $q \colon \mathcal{X} \to [-1, 1]$, we abuse notation and denote $q(\mathcal{D}) := \mathrm{E}_{x \sim \mathcal{D}}[q(x)]$, and similarly for $\mathcal{S} = (x_1, \ldots, x_n) \in \mathcal{X}^*$ we abuse notation and denote $q(\mathcal{S}) := \mathrm{E}_{x \leftarrow \mathcal{S}}[q(x)] = \frac{1}{n} \sum_{i=1}^{n} x_i$.

Fact 2.1 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables over $[-1, 1]$ with expectation μ. Then*

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \alpha\right] \leq 2 \cdot e^{-\alpha^2 n/2}$$

2.3 Cryptographic Primitives

2.3.1 Key Agreement Protocols

The most basic public-key cryptographic primitive is a (1-bit) *key agreement* protocol, defined below.

Definition 2.2 (key-agreement protocol)**.** *Let π be a two party protocol between two* interactive PPT *algorithms P_1 and P_2, each outputs 1-bit. Let $\pi(1^n)$ denote a random execution of the protocol on joint input 1^n (the security parameter), and let O_n^1, O_n^2 and T_n denote the random variables of P_1's output, P_2's output, and the transcript (respectively) in this execution. We say that π is an* (α, β)-key-agreement *protocol if the following holds for any* PPT *(i.e., "eavesdropper") A and every $n \in \mathbb{N}$:*

Agreement: $\Pr\left[O_n^1 = O_n^2\right] \geq \alpha(n)$*, and*

Secrecy: $\Pr\left[\mathsf{A}(T_n) = O_n^1\right] \leq \beta(n)$.

We say that π is a fully-secure key-agreement *protocol if it is an* $(1 - \text{neg}(n), 1/2 + \text{neg}(n))$-key-agreement *protocol.*

We use the following weaker type of agreement.

Definition 2.3 (Approximate agreement protocol)**.** *Let π be a two party protocol between two* interactive PPT *algorithms P_1 and P_2, each outputs a value in $[-1,1]$, and denote by $O_n^1, O_n^2 \in [-1,1]$ and T_n the random variables of the outputs of P_1, P_2 and transcript (respectively) in a random the execution $\pi(1^n)$. We say that π is an (α,β)-*approximate agreement *protocol if the following holds for any* PPT A *and $n \in \mathbb{N}$:*

Approximate Agreement: $\Pr\left[\left|O_n^1 - O_n^2\right| \leq \alpha(n)\right] \geq 1 - \mathrm{neg}(n)$*, and*

Secrecy: $\Pr\left[\left|\mathsf{A}(T_n) - O_n^1\right| \leq \beta(n)\right] \leq 1 - n^{-\Omega(1)}$.

Namely, when $\alpha(n) < \beta(n)$, the parties in an *approximate agreement* protocol do not agree on the same value, but are able to output values that are closer to each other than any prediction of a PPT "eavesdropper" adversary. We show that such approximate agreement suffices for constructing a fully-secure key-agreement.

Theorem 2.4. *Let $\alpha, \beta \colon \mathbb{N} \to [0,1]$ be efficiently computable functions such that $\alpha(n)/\beta(n) \leq n^{-\Omega(1)}$ and $\alpha(n) \cdot \beta(n) \geq 2^{-n}$ for large enough n. If there exists an (α,β)-approximate-agreement protocol, then there exists a fully-secure key-agreement protocol.*

A close variant of Theorem 2.4 is implicitly proved in [HMST22] (with better parameters, but in a more complicated setting). For completeness, we give a full proof of Theorem 2.4 in Appendix A.

2.3.2 Identity-Based Encryption

An Identity-Based Encryption (IBE) scheme [Sha84; Coc01; BF01] consists of four PPT algorithms $(\mathsf{Setup}, \mathsf{KeyGen}, \mathsf{Encrypt}, \mathsf{Decrypt})$ defined as follows:

- $\mathsf{Setup}(1^\lambda)$: given the security parameter λ, it outputs a master public key mpk and a master secret key msk.
- $\mathsf{KeyGen}(\mathsf{msk}, \mathsf{id})$: given the master secret key msk and an identity $\mathsf{id} \in [n]$, it outputs a decryption key $\mathsf{sk}_{\mathsf{id}}$.
- $\mathsf{Encrypt}(\mathsf{mpk}, \mathsf{id}, \mathsf{m})$: given the master public key mpk, and identity $\mathsf{id} \in [n]$ and a message m, it outputs a ciphertext ct.
- $\mathsf{Decrypt}(\mathsf{sk}_{\mathsf{id}}, \mathsf{ct})$: given a secret key $\mathsf{sk}_{\mathsf{id}}$ for identity id and a ciphertext ct, it outputs a string m.

The following are the properties of such an encryption scheme:

- **Completeness:** For all security parameter λ, identity $\mathsf{id} \in [n]$ and a message m, with probability 1 over $(\mathsf{mpk}, \mathsf{msk}) \sim \mathsf{Setup}(1^\lambda)$ and $\mathsf{sk}_{\mathsf{id}} \sim \mathsf{KeyGen}(\mathsf{msk}, \mathsf{id})$ it holds that

$$\mathsf{Decrypt}(\mathsf{sk}_{\mathsf{id}}, \mathsf{Encrypt}(\mathsf{mpk}, \mathsf{id}, \mathsf{m})) = \mathsf{m}$$

- **Security:** For any PPT adversary $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)$ it holds that:

$$\Pr[IND^{IBE}_{\mathsf{A}}(1^\lambda) = 1] \leq 1/2 + \mathrm{neg}(\lambda)$$

where IND^{IBE}_{A} is shown in Experiment 2.5.[5]

Experiment 2.5 ($IND^{IBE}_{\mathsf{A}}(1^\lambda)$)**.**

1. $(\mathsf{mpk}, \mathsf{msk}) \sim \mathsf{Setup}(1^\lambda)$.
2. $(\mathsf{id}^*, (\mathsf{m}^0_1, \ldots, \mathsf{m}^0_k), (\mathsf{m}^1_1, \ldots, \mathsf{m}^1_k), \mathsf{st}) \sim \mathsf{A}_1^{\mathsf{KeyGen}(\mathsf{msk},\cdot)}(\mathsf{mpk})$ *where* $|\mathsf{m}^0_i| = |\mathsf{m}^1_i|$ *for every* $i \in [k]$ *and for each query* id *by* A_1 *to* $\mathsf{KeyGen}(\mathsf{msk}, \cdot)$ *we have that* $\mathsf{id} \neq \mathsf{id}^*$.
3. *Sample* $b \leftarrow \{0, 1\}$.
4. *Sample* $\mathsf{ct}^*_i \sim \mathsf{Encrypt}(\mathsf{mpk}, \mathsf{id}^*, \mathsf{m}^b_i)$ *for every* $i \in [k]$.
5. $b' \sim \mathsf{A}_2^{\mathsf{KeyGen}(\mathsf{msk},\cdot)}(\mathsf{mpk}, (\mathsf{ct}^*_1, \ldots \mathsf{ct}^*_k), \mathsf{st})$ *where for each query* id *by* A_2 *to* $\mathsf{KeyGen}(\mathsf{msk}, \cdot)$ *we have that* $\mathsf{id} \neq \mathsf{id}^*$.
6. *Output* 1 *if* $b = b'$ *and* 0 *otherwise.*

Namely, the adversary chooses two sequences of messages $(\mathsf{m}^0_1, \ldots, \mathsf{m}^0_k)$ and $(\mathsf{m}^1_1, \ldots, \mathsf{m}^1_k)$, and gets encryptions of either the first sequence or the second one, where the encryptions made for identity id^* that the adversary does not hold its key (not allowed to query KeyGen on input id^*). The security requirement means that she cannot distinguish between the two cases (except with negligible probability).

Shamir [Sha84] was the first to consider the problem of constructing an IBE scheme that can support many identities using small keys. The first IBE schemes were realized by Boneh and Franklin [BF01] and Cocks [Coc01], but their security analyses were based on non-standard cryptographic assumptions: the quadratic residuocity assumption [Coc01] and assumptions on groups with billinear map [BF01]. More recently, Döttling and Garg [DG21] and Blazy and Kakvi [BK22] have managed to construct an IBE scheme based on the standard Computational Diffie-Hellman (CDH) hardness assumption. Below we summarize the construction properties of [DG21].

Theorem 2.6 ([DG21])**.** *Assume that the Computational Diffie-Hellman (CDH) Problem is hard. Then there exists an IBE scheme* $\mathcal{E} = (\mathsf{Setup}, \mathsf{KeyGen}, \mathsf{Encrypt}, \mathsf{Decrypt})$ *for* n *identities such that given a security parameter* λ*, the master keys and each decryption key are of size* $O(\lambda \cdot \log n)$.[6]

2.4 Balanced Adaptive Data Analysis

Adaptive data analysis is modeled as a game between a *mechanism* M and an *analyst* A. The mechanism gets as input n i.i.d. samples $x_1, \ldots, x_n$ from an (unknown) distribution $\mathcal{D}$ over a domain $\mathcal{X}$, and its goal is to answer statistical queries about $\mathcal{D}$, produced by the analyst. Namely, when A sends a statistical query $q\colon \mathcal{X} \to [-1, 1]$, M is required to return an answer $y \in [-1, 1]$ that is close to $q(\mathcal{D}) = \mathrm{E}_{x \sim \mathcal{D}}[q(x)]$.

[5] The IBE security experiment is usually described as Experiment 2.5 with $k = 1$ (i.e., encrypting a single message). Yet, it can be extended to any sequence of messages using a simple reduction.

[6] The construction can also be based on the hardness of *factoring*.

In this work we investigate a *balanced* setting where the adversarial analyst does not have an informational advantage over the mechanism. Namely, the analyst has no knowledge about the underline distribution which the mechanism does not have. We model this situation by separating between the analyst from the underline distribution, as follows:

The mechanism plays a game with a *balanced* adversary that consists of two (isolated) algorithms: a *sampler* A_1, which chooses a distribution $\mathcal{D}$ over a domain $\mathcal{X}$ and provides n i.i.d. samples to M, and an *analyst* A_2, which asks the adaptive queries about the distribution (see Game 2.7). The public inputs for the ADA game are the number of samples n, the number of queries ℓ and the domain $\mathcal{X}$. Since this work mainly deals with computationally bounded algorithms that we would like to model as PPT algorithms, we provide n and ℓ in unary representation. We also assume for simplicity that $\mathcal{X}$ is finite, which allows to represent each element as a binary vector of dimension $\lceil \log|\mathcal{X}| \rceil$, and we provide the dimension in unary representation as well.

Game 2.7 ($\mathsf{ADA}_{n,\ell,\mathcal{X}}[\mathsf{M}, \mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)]$, redefinition of Game 1.3)**.**

Public inputs: Number of samples 1^n*, number of queries* 1^ℓ*, and a domain* $\mathcal{X}$ *(represented as* $1^{\lceil \log|\mathcal{X}| \rceil}$*).*

Operation:

1. A_1 *chooses a distribution* $\mathcal{D}$ *over* $\mathcal{X}$ *(specified by a sampling algorithm) and sends* $\mathcal{S} = (x_1, \ldots, x_n) \sim \mathcal{D}^n$ *to* M *(i.e., applies the sampling algorithm* n *times).*
2. *For* $i = 1, \ldots, \ell$*:*
   (a) A_2 *sends a query* $q_i \colon \mathcal{X} \mapsto [-1, 1]$ *to* M*.*
   (b) M *sends an answer* $y_i \in [-1, 1]$ *to* A_2*.*
   (As A_2 *and* M *are stateful,* q_i *and* y_i *may depend on the previous messages.)*
3. *The outcome is one if* $\exists i \in [\ell]$ *s.t.* $|y_i - q_i(\mathcal{D})| > 1/10$*, and zero otherwise.*

All previous negative results ([HU14; SU15b; DSWZ23]) where achieved by reduction to a restricted family of mechanisms, called *natural* mechanisms (Definition 1.8). These are algorithms that can only evaluate the query on the sample points they are given.

For *natural* mechanisms (even unbounded ones), the following was proven.

Theorem 2.8 ([HU14; SU15b])**.** *There exists a pair of* PPT *algorithms* $\widetilde{\mathsf{A}} = (\widetilde{\mathsf{A}}_1, \widetilde{\mathsf{A}}_2)$ *such that for every* natural *mechanism* $\widetilde{\mathsf{M}}$ *and every large enough* n *and* $\ell = \Theta(n^2)$ *it holds that*

$$\Pr\left[\mathsf{ADA}_{n,\ell,\mathcal{X}=[2000n]}[\widetilde{\mathsf{M}}, \widetilde{\mathsf{A}}] = 1\right] > 3/4. \tag{1}$$

In particular, $\widetilde{\mathsf{A}}_1$ *always chooses the uniform distribution over* $[2000n]$*, and* $\widetilde{\mathsf{A}}_2$ *uses only queries over the range* $\{-1, 0, 1\}$*.*

The adversary $\widetilde{\mathsf{A}}$ from Theorem 2.8 uses queries that are based on random *interactive fingerprinting code* [SU15b] which enables to reconstruct most of the n samples $\mathcal{S}$ given $\Theta(n^2)$ accurate answers that are only a function of the samples (as a *natural* mechanism must behave). Once

the samples are revealed to the analyst, it then prepares a last query that cannot be answered accurately by a *natural* mechanism (e.g., a query q with $q(x) = 0$ for $x \in \mathcal{S}$, but with different values for elements $x \in \mathcal{X} \setminus \mathcal{S}$). In particular, this holds for the mechanism $\widetilde{\mathsf{M}}$ which given a query q_i for $i \in [\ell - 1]$ and a sample $\mathcal{S}$, answers the empirical mean $q(\mathcal{S}) = \frac{1}{n}\sum_{x\in\mathcal{S}} x$. See the observation below which is used in Section 4.

Observation 2.9 (Implicit in [SU15b])**.** *Let $\widetilde{\mathsf{M}}$ be a* natural *mechanism that given a sample $\mathcal{S} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and a query $q\colon \mathcal{X} \to [-1, 1]$ which is not the last one, answers the empirical mean $q(\mathcal{S}) = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then in a random execution of $\mathsf{ADA}_{n,\ell,\mathcal{X}}[\widetilde{\mathsf{M}}, \widetilde{\mathsf{A}}]$ ($\widetilde{\mathsf{A}}, n, \ell, \mathcal{X}$ as in Theorem 2.8), $\widetilde{\mathsf{M}}$ will fail to answer the last query accurately, regardless of what natural strategy it uses for this query.*

3 Constructing a Balanced Adversary via IBE

In this section we prove that, under standard public-key cryptography assumptions (in particular, the existence of an IBE scheme), there is an efficient reduction to *natural* mechanisms that holds against any PPT mechanism, yielding a general lower bound for the computational case. In particular, we show that there exists a pair of PPT algorithms $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)$ such that for every PPT mechanism M it holds that

$$\Pr\left[\mathsf{ADA}_{n,\ell=\Theta(n^2),\mathcal{X}=\{0,1\}^{\widetilde{O}(\lambda)}}[\mathsf{M}, \mathsf{A}] = 1\right] \geq 3/4 - \mathrm{neg}(n),$$

for any function $\lambda = \lambda(n)$ such that $n \leq \mathrm{poly}(\lambda)$ (e.g., $\lambda = n^{0.1}$). More specifically, we use a domain $\mathcal{X}$ with $\log|\mathcal{X}| = 2k + \log n + O(1)$, where $k = k(n)$ is the keys' length in the IBE scheme with security parameter λ that supports $O(n)$ identities (by Theorem 2.6, such an IBE scheme exists with $k = O(\lambda \cdot \log n)$ under the CDH hardness assumption). A_1 is defined in Algorithm 3.1, and A_2 is defined in Algorithm 3.2.

Algorithm 3.1 (Sampler A_1)**.**

Inputs: *Number of samples 1^n, number of queries 1^ℓ and domain $\mathcal{X}$ (defined below). Let $m = 2000n$.*

Oracle Access: A_1 *has access to an IBE scheme $\mathcal{E} = (\mathsf{Setup}, \mathsf{KeyGen}, \mathsf{Encrypt}, \mathsf{Decrypt})$ that supports m identities with security parameter $\lambda = \lambda(n)$. Let $k = k(n)$ be the sizes of the keys in this scheme. Let $\mathcal{X} = [m] \times \{0,1\}^{2k}$.*

Setting: A_1 *is the* sampler *in the $\mathsf{ADA}_{n,\ell,\mathcal{X}}$ game (Game 2.7) which provides to M n i.i.d. samples from some underline distribution $\mathcal{D}$.*

Operation: % Step 1 of Game 2.7:

- *Sample $(\mathsf{mpk}, \mathsf{msk}) \sim \mathsf{Setup}(1^\lambda)$ and $\mathsf{sk}_j \sim \mathsf{KeyGen}(\mathsf{msk}, j)$ for every $j \in [m]$, and let $\mathcal{T} = \{(j, \mathsf{mpk}, \mathsf{sk}_j)\}_{j=1}^{m}$ and $\mathcal{D} = \mathcal{U}_\mathcal{T}$ (i.e., the uniform distribution over the triplets in $\mathcal{T}$).*
- *Send to M n i.i.d. samples $\mathcal{S} \sim \mathcal{D}^n$.*

Algorithm 3.2 (Analyst A_2)**.**

Inputs: *Number of samples* 1^n*, number of queries* 1^ℓ *and a domain* $\mathcal{X}$ *(defined below). Let* $m = 2000n$*.*

Oracle access: A_2 *has access to an IBE scheme* $\mathcal{E} = (\mathsf{Setup}, \mathsf{KeyGen}, \mathsf{Encrypt}, \mathsf{Decrypt})$ *that supports* m *identities with security parameter* $\lambda = \lambda(n)$*. Let* $k = k(n)$ *be the sizes of the keys in this scheme. Let* $\mathcal{X} = [m] \times \{0,1\}^{2k}$*.*

Setting: A_2 *is the* analyst *in the* $\mathsf{ADA}_{n,\ell,\mathcal{X}}$ *game (Game 2.7). It has access to the analyst* $\widetilde{\mathsf{A}}_2$ *from Theorem 2.8 and it interacts with a (general, not necessarily natural) mechanism* M *in* $\mathsf{ADA}_{n,\ell,\mathcal{X}}[\mathsf{M}, (\mathsf{A}_1, \cdot)]$ *where* A_1 *is Algorithm 3.1.*

Operation:

1. *% The first k iterations in Step 2 of Game 2.7:*

   For $i = 1, 2, \ldots, k$*:*

   (a) *% Step 2a: Send to* M *a query* q_i *that on input* $(j, x, y) \in [m] \times \{0,1\}^k \times \{0,1\}^k$ *outputs* x_i*.*

   (b) *% Step 2b: Receive an answer from* M *and round it for reconstructing the* i^{th} *bit of* mpk*.*

2. *Initialize an emulation of* $\widetilde{\mathsf{A}}_2$ *in the game* $\mathsf{ADA}_{n,\ell-k,[m]}$*.*

3. *% The last* $\ell - k$ *iterations in Step 2 of Game 2.7:*

   For $i = 1, \ldots, \ell - k$*:*

   (a) *Obtain the* i^{th} *query* $\tilde{q}_i \colon [m] \to \{-1, 0, 1\}$ *of the emulated* $\widetilde{\mathsf{A}}_2$*.*

   (b) *For* $j \in [m]$*, compute* $\mathsf{ct}_{i,j} = \mathsf{Encrypt}(\mathsf{mpk}, j, \tilde{q}_i(j))$ *(i.e., encrypt* $\tilde{q}_i(j)$ *for identity* j*).*

   (c) *Define the query* $q_{i+k} \colon \mathcal{X} \to \{-1, 0, 1\}$ *that on input* $(j, x, y) \in [m] \times \{0,1\}^k \times \{0,1\}^k$ *outputs* $\mathsf{Decrypt}(y, \mathsf{ct}_{i,j})$*. The description of* q_{i+k} *consists of* $\{\mathsf{ct}_{i,j}\}_{j \in [m]}$*.*

   (d) *% Step 2a: Send (the description of)* q_{i+k} *to* M*.*

   (e) *% Step 2b: Receive an answer* y_{i+k} *from* M*.*

   (f) *Send* $\tilde{y}_i = y_{i+k}$ *to the emulated* $\widetilde{\mathsf{A}}_2$ *(as an answer to* $\tilde{q}_i$*).*

Theorem 3.3 (Restatement of Theorem 1.5)**.** *Assume the existence of an IBE scheme* $\mathcal{E}$ *that supports* $m = 2000n$ *identities with security parameter* $\lambda = \lambda(n)$ *s.t.* $n \leq \text{poly}(\lambda)$ *using keys of length* $k = k(n)$*. Let* $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)$*, where* A_1 *is Algorithm 3.1 and* A_2 *is Algorithm 3.2. Then there exists* $\ell = \Theta(n^2) + k$ *such that for every* PPT *mechanism* M *it holds that*

$$\Pr\left[\mathsf{ADA}_{n,\ell,\mathcal{X}=[m]\times\{0,1\}^{2k}}[\mathsf{M}, \mathsf{A}] = 1\right] > 3/4 - \text{neg}(n).$$

Proof. Fix a PPT mechanism M and large enough n. Consider the mechanism $\widetilde{\mathsf{M}}$ defined in Algorithm 3.4 with respect to M. First, note that $\widetilde{\mathsf{M}}$ is indeed *natural* since, upon receiving the query

$\tilde{q}_i$, it does not use the values $\{\tilde{q}_i(j)\}_{j\in[m]\setminus\mathcal{J}}$. Therefore, by Theorem 2.8 it holds that

$$\Pr\Big[\mathsf{ADA}_{n,\ell-k,[m]}[\widetilde{\mathsf{M}},\widetilde{\mathsf{A}}]=1\Big]\geq 3/4.$$

In the following, let $\widetilde{\mathsf{M}}'$ be an (unnatural) variant of $\widetilde{\mathsf{M}}$ that operates almost the same, except that in Step 4b, rather than sampling $\mathsf{ct}_{i,j}\sim\mathsf{Encrypt}(\mathsf{mpk},j,0)$ for $j\notin\mathcal{J}$, it samples $\mathsf{ct}_{i,j}\sim\mathsf{Encrypt}(\mathsf{mpk},j,\tilde{q}_i(j))$. Note that both $\widetilde{\mathsf{M}}$ and $\widetilde{\mathsf{M}}'$, when playing in $\mathsf{ADA}_{n,\ell-k,[m]}[\cdot,\widetilde{\mathsf{A}}]$, emulate an execution of $\mathsf{ADA}_{n,\ell,\mathcal{X}}[\mathsf{M},\mathsf{A}]$, where the only difference between them is the values of $\mathsf{ct}_{i,j}$ for $j\notin\mathcal{J}$ that they send to the emulated M in each iteration i. But M is a $\mathrm{poly}(\lambda)$-time mechanism and its view in the emultations does not contain the keys $\{\mathsf{sk}_j\}_{j\notin\mathcal{J}}$. Therefore, by the security guarantee of the IBE scheme, the behavior of the emulated M is indistinguishable in both executions, yielding that

$$\begin{aligned}\Pr\Big[\mathsf{ADA}_{n,\ell-k,[m]}[\widetilde{\mathsf{M}}',\widetilde{\mathsf{A}}]=1\Big]&\geq\Pr\Big[\mathsf{ADA}_{n,\ell-k,[m]}[\widetilde{\mathsf{M}},\widetilde{\mathsf{A}}]=1\Big]-\mathrm{neg}(n)\\&\geq 3/4-\mathrm{neg}(n).\end{aligned}$$

In the following we focus on proving that

$$\Pr\big[\mathsf{ADA}_{n,\ell,\mathcal{X}}[\mathsf{M},\mathsf{A}]=1\big]\geq\Pr\Big[\mathsf{ADA}_{n,\ell-k,[m]}[\widetilde{\mathsf{M}}',\widetilde{\mathsf{A}}]=1\Big],\tag{2}$$

which concludes the proof.

Let $T,\ Q_1,Y_1,\ldots,Q_\ell,Y_\ell$ be the (r.v.'s of the) values of $\mathcal{T},\ q_1,y_1,\ldots,q_\ell,y_\ell$ (respectively) induced by A_2 in a random execution of $\mathsf{ADA}_{n,\ell,\mathcal{X}}[\mathsf{M},\mathsf{A}]$, and let E be the event that $\forall i\in[k]:|Q_i(\mathcal{U}_T)-Y_i|\leq 1/10$. Similarly, let $T',\ Q'_1,Y'_1,\ldots,Q'_\ell,Y'_\ell,\ \widetilde{Q}_1,\widetilde{Y}_1,\ldots,\widetilde{Q}_{\ell-k},\widetilde{Y}_{\ell-k}$ be the (r.v.'s of the) values of $\mathcal{T},\ q_1,y_1,\ldots,q_\ell,y_\ell,\ \tilde{q}_1,\tilde{y}_1,\ldots,\tilde{q}_{\ell-k},\tilde{y}_{\ell-k}$ induced by $\widetilde{\mathsf{M}}'$ in an (independent) execution of $\mathsf{ADA}_{n,\ell-k,[m]}[\widetilde{\mathsf{M}}',\widetilde{\mathsf{A}}]$, and let E' be the event that $\forall i\in[k]:|Q'_i(\mathcal{U}_{T'})-Y'_i|\leq 1/10$. By construction, the following holds:

1. $\Pr[E]=\Pr[E']$ (holds since $\widetilde{\mathsf{M}}'$, as $\widetilde{\mathsf{M}}$, perfectly emulates the first k queries and answers of $\mathsf{ADA}_{n,\ell,\mathcal{X}}[\mathsf{M},\mathsf{A}]$ in Step 3b),

2. $(Q_i(\mathcal{U}_T),Y_i)_{i=k+1}^{\ell}|_E\equiv(Q'_i(\mathcal{U}_{T'}),Y'_i)_{i=k+1}^{\ell}|_{E'}$ (Conditioned on E, A_2 successfully reconstruct the master public key mpk in Step 1b. This yields that conditioned on E' in $\mathsf{ADA}_{n,\ell-k,[m]}[\widetilde{\mathsf{M}}',\widetilde{\mathsf{A}}]$, $\widetilde{\mathsf{M}}'$ perfectly emulates an execution of $\mathsf{ADA}_{n,\ell,\mathcal{X}}[\mathsf{M},\mathsf{A}]$ conditioned on E), and

3. $(Q'_i(\mathcal{U}_{T'}),Y'_i)_{i=k+1}^{\ell}\equiv(\widetilde{Q}_i(\mathcal{U}_{[m]}),\widetilde{Y}_i)_{i=1}^{\ell-k}$ ($\widetilde{\mathsf{M}}'$ defines each q_{i+k} by encrypting all the outputs of $\tilde{q}_i$, so in the execution of $\widetilde{\mathsf{M}}'$, for every $i\in[\ell-k]$ it always holds that $q_{i+k}(\mathcal{U}_{\mathcal{T}})=\tilde{q}_i(\mathcal{U}_{[m]})$, and it also holds that $y_{i+k}=\tilde{y}_i$ by Step 4e).

Hence, we conclude that

$$
\begin{aligned}
&\Pr\left[\mathsf{ADA}_{n,\ell,\mathcal{X}}[\mathsf{M},\mathsf{A}]=1\right]\\
&=\Pr[\exists i\in[\ell] \text{ s.t. } |Q_i(\mathcal{U}_T)-Y_i|>1/10]\\
&=\Pr[\exists i\in\{k+1,\ldots,\ell\} \text{ s.t. } |Q_i(\mathcal{U}_T)-Y_i|>1/10 \mid E]\cdot\Pr[E]+1\cdot\Pr[\neg E]\\
&=\Pr\left[\exists i\in\{k+1,\ldots,\ell\} \text{ s.t. } \left|Q'_i(\mathcal{U}_{T'})-Y'_i\right|>1/10 \mid E'\right]\cdot\Pr\left[E'\right]+\Pr\left[\neg E'\right]\\
&\geq\Pr\left[\exists i\in\{k+1,\ldots,\ell\} \text{ s.t. } \left|Q'_i(\mathcal{U}_{T'})-Y'_i\right|>1/10\right]\\
&=\Pr\left[\exists i\in[\ell-k] \text{ s.t. } \left|\widetilde{Q}_i(U_{[m]})-\widetilde{Y}_i\right|>1/10\right]\\
&=\Pr\left[\mathsf{ADA}_{n,\ell-k,[m]}[\widetilde{\mathsf{M}}',\widetilde{\mathsf{A}}]=1\right],
\end{aligned}
$$

as required. The third equality holds by Items 1 and 2 and the penultimate one holds by Item 3. □

Algorithm 3.4 (Natural mechanism $\widetilde{\mathsf{M}}$).

Public parameters: *Number of samples* 1^n*, number of queries* 1^ℓ*, and a domain* $\widetilde{\mathcal{X}} = [m]$ *for* $m = 2000n$*.*

Oracle Access: $\widetilde{\mathsf{M}}$ *has access to an IBE scheme* $\mathcal{E} = (\mathsf{Setup}, \mathsf{KeyGen}, \mathsf{Encrypt}, \mathsf{Decrypt})$ *that supports* m *identities with security parameter* $\lambda = \lambda(n)$*. Let* $k = k(n)$ *be the sizes of the keys in this scheme.*

Setting: $\widetilde{\mathsf{M}}$ *has access to a mechanism* M *and to algorithms* A_1 *and* A_2 *(Algorithms 3.1 and 3.2, respectively) and interacts in* $\mathsf{ADA}_{n,\ell,[m]}[\cdot, \widetilde{\mathsf{A}}]$ *(Game 2.7), where* $\widetilde{\mathsf{A}} = (\widetilde{\mathsf{A}}_1, \widetilde{\mathsf{A}}_2)$ *is the pair of algorithms from Theorem 2.8.*

Operation:

1. *% Step 1 of Game 2.7: Receive* $\mathcal{J} \leftarrow [m]^n$ *from* $\widetilde{\mathsf{A}}_1$*.*
2. *Sample* $(\mathsf{mpk}, \mathsf{msk}) \sim \mathsf{Setup}(1^\lambda)$ *and* $\mathsf{sk}_j \sim \mathsf{KeyGen}(\mathsf{msk}, j)$ *for each* $j \in [m]$*.*
3. *Start an emulation of* M *in the game* $\mathsf{ADA}_{n,\ell+k,\mathcal{X}}[\cdot, \mathsf{A}]$ *for* $\mathcal{X} = [m] \times \{0,1\}^{2k}$*, where:*
   (a) *In Step 1 of the emulation, let* M *receive the samples* $\mathcal{S} = \{(j, \mathsf{mpk}, \mathsf{sk}_j)\}_{j \in \mathcal{J}}$ *which plays the role of* n *i.i.d. samples from* $\mathcal{T} = \{(j, \mathsf{mpk}, \mathsf{sk}_j)\}_{j \in [m]}$ *(i.e., the* n *samples that* A_1 *sends to* M *in the emulation).*
   (b) *Emulate the first* k *queries and answers* $q_1, y_1, \ldots, q_k, y_k$ *when interacting with* A_2 *in* $\mathsf{ADA}_{n,\ell+k,\mathcal{X}}[\mathsf{M}, \mathsf{A}]$ *(Step 1 of Algorithm 3.2).*
4. *% Step 2 of Game 2.7:*
   For $i = 1, \ldots, \ell$*:*
   (a) *% Step 2a: Receive a query* $\tilde{q}_i \colon [m] \to \{-1, 0, 1\}$ *from* $\widetilde{\mathsf{A}}_2$*.*
   (b) *For* $j \in \mathcal{J}$ *compute* $\mathsf{ct}_{i,j} \sim \mathsf{Encrypt}(\mathsf{mpk}, j, \tilde{q}_i(j))$ *and for* $j \in [m] \setminus \mathcal{J}$ *compute* $\mathsf{ct}_{i,j} \sim \mathsf{Encrypt}(\mathsf{mpk}, j, 0)$*.*
   (c) *Continue the emulation of* $\mathsf{ADA}_{n,\ell+k,\mathcal{X}}[\mathsf{M}, \mathsf{A}]$ *by sending* $\{\mathsf{ct}_{i,j}\}_{j=1}^m$ *to* M *as the* $(k+i)$*'th query* q_{i+k} *of* A_2*.*
   (d) *Let* y_{k+i} *be the answer that* M *sends in the emulation (in response to the* $(k+i)$*'th query).*
   (e) *% Step 2b: Send the answer* $\tilde{y}_i = y_{k+i}$ *to* $\widetilde{\mathsf{A}}_2$*.*

4 Reduction to Natural Mechanisms Implies Key Agreement

In this section we prove that any PPT *balanced* adversary $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)$ that has the structure of all known lower bounds ([HU14; SU15b; DSWZ23] and ours in Section 3), can be used to construct a *key-agreement* protocol.

All known constructions use an adversary A that wraps the adversary $\widetilde{\mathsf{A}}$ for the natural mecha-

nisms case (Theorem 2.8) by forcing every mechanism M to behave *naturally* using cryptography. In particular, they all have the following two key properties that are inherited from $\widetilde{\mathsf{A}}$:

1. The analyst asks queries that it knows the true answer to them (i.e., the true answer can be extracted from its view), and
2. If a PPT mechanism attempts to behave accurately (e.g., given a query, it answers the empirical mean), then in the last round it will fail with high probability (which is the analog of Observation 2.9).

The formal statement is given in the following theorem.

Theorem 4.1 (Restatement of Theorem 1.6)**.** *Assume the existence of a* PPT *adversary* $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)$ *and functions* $\ell = \ell(n) \leq \text{poly}(n)$ *and* $\mathcal{X} = \mathcal{X}(n)$ *with* $\log|\mathcal{X}| \leq \text{poly}(n)$ *such that the following holds: Let* $n \in \mathbb{N}$ *and consider a random execution of* $\mathsf{ADA}_{n,\ell,\mathcal{X}}[\mathsf{M}, \mathsf{A}]$ *where* M *is the mechanism that given a sample* $\mathcal{S}$ *and a query* q*, answers the empirical mean* $q(\mathcal{S})$*. Let* D_n *and* Q_n *be the (r.v.'s of the) values of* $\mathcal{D}$ *and* $q = q_\ell$ *(the last query) in the execution (respectively), let* T_n *be the transcript of the execution between the analyst* A_2 *and the mechanism* M *(i.e., the queries and answers), and let* V_n *be the view of* A_2 *at the end of the execution (without loss of generality, its input, random coins and the transcript). Assume that*

1. $\exists$ PPT *algorithm* F *s.t.* $\forall n \in \mathbb{N}: \Pr\left[|\mathsf{F}(V_n) - Q_n(D_n)| \leq n^{-1/10}\right] \geq 1 - \text{neg}(n)$*, and*
2. $\forall$ PPT *algorithm* G *and* $\forall n \in \mathbb{N}: \Pr[|\mathsf{G}(T_n) - Q_n(D_n)| \leq 1/10] \leq 1/4 + \text{neg}(n)$.

Then using A *and* F *it is possible to construct a fully-secure key-agreement protocol.*

Note that Assumption 1 in Theorem 4.1 formalizes the first property in which the analyst knows a good estimation of the true answer, and the PPT algorithm F is the assumed knowledge extractor. Assumption 2 in Theorem 4.1 formalizes the second property which states that the mechanism, which answers the empirical mean along the way, will fail in the last query, no matter how it chooses to act (this behavior is captured with the PPT algorithm G), and moreover, it is enough to assume that this requirement only holds with respect to to transcript of the execution, and not with respect to the view of the mechanism.

Example: Our Adversary from Section 3 In the following we show that our adversary $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)$ (Algorithms 3.1 and 3.2) has the above properties. Using similar arguments it can be shown that any previously known lower bound [HU14; SU15b; DSWZ23] has these properties as well.

Recall that the *sampler* A_1 first samples keys $\mathsf{mpk}, \mathsf{msk}, \{\mathsf{sk}_j\}_{j=1}^m$ ($m = 2000n$), and generates n uniformly random samples from the triplets $\mathcal{T} = \{(j, \mathsf{mpk}, \mathsf{sk}_j)\}_{j=1}^m$. The mechanism M gets the samples, and assume that M simply outputs the empirical mean of each query. Therefore, in the first k queries of the interaction, the analyst A_2 discovers the master key mpk. This allows it to wrap each query $\tilde{q}_i$ of the analyst $\widetilde{\mathsf{A}}_2$ by encrypting all the values using mpk, and sending the encryptions to M. Therefore, it is clear that A_2 knows the true mean for each wrapped query q_i (and in particular, the last one), since it equals to $\frac{1}{m}\sum_{j=1}^m \tilde{q}_i(j)$ (a description of $\tilde{q}_i$ is part of the view of A_2). This fulfills Assumption 1 of Theorem 4.1.

Regarding Assumption 2, note that when M simply answers the empirical means along the way, it is translated to answering the empirical means to the analyst $\widetilde{\mathsf{A}}_2$. Therefore, by Observation 2.9, $\widetilde{\mathsf{A}}_2$ will provide a last query that fails each last round strategy. By the properties of the IBE scheme, M does not see any values beyond its n samples, which forces it to behave like a *natural* mechanism. In particular, the above also holds when M uses a last round strategy G which is only a function of the transcript (which is only part of the view of M).

4.1 Proving Theorem 4.1

Theorem 4.1 is an immediate corollary of the following Lemma 4.3 and Theorem 2.4.

Protocol 4.2 (Approximate agreement protocol $(\mathsf{P}_1, \mathsf{P}_2)$)**.**

Input: *A security parameter 1^n. Let $\ell = \ell(n)$ and $\mathcal{X} = \mathcal{X}(n)$ be as in Theorem 4.1.*

Access: *Each P_i , for $i \in [2]$, has access to algorithm A_i from Theorem 4.1. P_2 has also access to algorithm F from Theorem 4.1.*

Operation:

- P_1 *emulates A_1 on input 1^n, 1^ℓ, and $\mathcal{X}$ for obtaining a distribution $\mathcal{D}$ (specified by a sampling procedure), and then samples $2n$ i.i.d. samples according to it. Let $\mathcal{S}$ be the first n samples, and let $\mathcal{S}'$ be the last n samples.*
- P_2 *initializes an emulation of A_2 on inputs 1^n, 1^ℓ, and $\mathcal{X}$.*
- *For $i = 1$ to ℓ:*
  1. P_2 *receive the i^{th} query q_i from the emulated A_2 and send it to P_1.*
  2. P_1 *sends $y_i = q_i(\mathcal{S})$ to P_2.*
  3. P_2 *sends y_i as the i^{th} answer to the emulated A_2.*
- P_1 *outputs $q_\ell(\mathcal{S}')$.*
- P_2 *outputs $\mathsf{F}(v)$ where v is the view of A_2 in the emulation.*

Lemma 4.3. *Let $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)$, $\ell = \ell(n)$, $\mathcal{X} = \mathcal{X}(n)$, and F be as in Theorem 4.1. Then Protocol 4.2 (w.r.t. these values) is an* $(2 \cdot n^{-1/10}, 1/20)$-approximate agreement *protocol according to Definition 2.3.*

Proof. Consider a random execution of $(\mathsf{P}_1, \mathsf{P}_2)$ on input 1^n. Let D_n, S_n, S'_n, Q_n, V_n be the values of $\mathcal{D}, \mathcal{S}, \mathcal{S}', q_\ell, v$ (respectively), and let $O_n^1 = Q(S'_n)$ and $O_n^2 = \mathsf{F}(V)$ be the outputs of P_1 and P_2 (respectively). Let E be the event $|Q_n(S'_n) - Q(D_n)| \leq n^{-1/10}$. Note that Q_n and S'_n are independent conditioned on D_n, and S'_n contains n i.i.d. samples from D_n. Hence by Hoeffding's inequality (Fact 2.1) it holds that $\Pr[E] \geq 1 - \text{neg}(n)$.

The agreement guarantee holds by the following computation.

$$\begin{aligned}
\Pr\left[\left|O_n^1 - O_n^2\right| \le 2 \cdot n^{-1/10}\right] &= \Pr\left[\left|Q_n(S'_n) - \mathsf{F}(V)\right| \le 2 \cdot n^{-1/10}\right] \\
&\ge \Pr\left[\left|Q_n(S'_n) - \mathsf{F}(V)\right| \le 2 \cdot n^{-1/10} \mid E\right] \cdot \Pr[E] \\
&\ge \Pr\left[\left|Q_n(D_n) - \mathsf{F}(V)\right| \le n^{-1/10} \mid E\right] \cdot \Pr[E] \\
&\ge \Pr\left[\left|Q_n(D_n) - \mathsf{F}(V)\right| \le n^{-1/10}\right] - \Pr[\neg E] \\
&\ge 3/4 - \mathrm{neg}(n),
\end{aligned}$$

where the last inequality holds by Assumption 1.

For the secrecy guarantee, note that the transcript T_n between P_1 and P_2 only consists of the transcript between the analysis and the mechanism in $\mathsf{ADA}_{n,\ell,\mathcal{X}}[\mathsf{M},\mathsf{A}]$ where M is the mechanism that answers the empirical mean of each query (holds by the answers that P_1 sends in Step 2). Therefore, for every PPT adversary G we conclude that

$$\begin{aligned}
\Pr\left[\left|G(T_n) - O_n^1\right| \le 1/20\right] &\le \Pr\left[\left|G(T_n) - Q_n(S'_n)\right| \le 1/20 \mid E\right] \cdot \Pr[E] + \Pr[\neg E] \\
&\le \Pr[|G(T_n) - Q_n(D_n)| \le 1/10 \mid E] \cdot \Pr[E] + \Pr[\neg E] \\
&\le \Pr[|G(T_n) - Q_n(D_n)| \le 1/10] + \Pr[\neg E] \\
&\le 1/4 + \mathrm{neg}(n),
\end{aligned}$$

as required. The last inequality holds by Assumption 2.

□

References

[BF01] D. Boneh and M. K. Franklin, "Identity-based encryption from the weil pairing," in *Advances in Cryptology - CRYPTO 2001, 21st Annual International Cryptology Conference, Proceedings*, J. Kilian, Ed., vol. 2139, 2001, pp. 213–229 (cit. on pp. 5, 10, 11).

[BK22] O. Blazy and S. A. Kakvi, "Identity-based encryption in DDH hard groups," in *Proceedings of the 13th International Conference on Cryptology in Africa, AFRICACRYPT 2022*, 2022, pp. 81–102 (cit. on p. 11).

[BKM+22] A. Beimel, H. Kaplan, Y. Mansour, K. Nissim, T. Saranurak, and U. Stemmer, "Dynamic algorithms against an adaptive adversary: Generic constructions and lower bounds," in *STOC*, ACM, 2022, pp. 1671–1684 (cit. on p. 8).

[Bla23] G. Blanc, "Subsampling suffices for adaptive data analysis," *CoRR*, vol. abs/2302.08661, 2023 (cit. on pp. 1, 2).

[BNS+16] R. Bassily, K. Nissim, A. D. Smith, T. Steinke, U. Stemmer, and J. Ullman, "Algorithmic stability for adaptive data analysis," in *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016*, 2016, pp. 1046–1059 (cit. on pp. 1, 2).

[Coc01] C. C. Cocks, "An identity based encryption scheme based on quadratic residues," in *Cryptography and Coding, 8th IMA International Conference, Proceedings*, vol. 2260, 2001, pp. 360–363 (cit. on pp. 5, 10, 11).

[DFH+15a] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "Generalization in adaptive data analysis and holdout reuse," in *Advances in Neural Information Processing Systems (NIPS)*, 2015 (cit. on pp. 1, 2).

[DFH+15b] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "The reusable holdout: Preserving validity in adaptive data analysis," *Science*, vol. 349, no. 6248, pp. 636–638, 2015 (cit. on pp. 1, 2).

[DFH+15c] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth, "Preserving statistical validity in adaptive data analysis," in *STOC*, ACM, 2015, pp. 117–126 (cit. on pp. 1, 2).

[DG21] N. Döttling and S. Garg, "Identity-based encryption from the diffie-hellman assumption," *J. ACM*, vol. 68, no. 3, 14:1–14:46, 2021 (cit. on pp. 6, 11).

[DH76] W. Diffie and M. E. Hellman, "New directions in cryptography," *IEEE Trans. Inf. Theory*, vol. 22, no. 6, pp. 644–654, 1976 (cit. on p. 6).

[DK22] Y. Dagan and G. Kur, "A bounded-noise mechanism for differential privacy," in *Conference on Learning Theory*, vol. 178, PMLR, 2022, pp. 625–661 (cit. on pp. 1, 2).

[DSWZ23] I. Dinur, U. Stemmer, D. P. Woodruff, and S. Zhou, "On differential privacy and adaptive data analysis with bounded space," *CoRR*, vol. abs/2302.05707, 2023 (cit. on pp. 1, 8, 12, 17, 18).

[Eld16] S. Elder, "Challenges in bayesian adaptive data analysis," *CoRR*, vol. abs/1604.02492, 2016. [Online]. Available: http://arxiv.org/abs/1604.02492 (cit. on pp. 4, 5).

[FRR20] B. Fish, L. Reyzin, and B. I. P. Rubinstein, "Sampling without compromising accuracy in adaptive data analysis," in *Algorithmic Learning Theory, ALT 2020*, vol. 117, PMLR, 2020, pp. 297–318 (cit. on p. 1).

[FS17] V. Feldman and T. Steinke, "Generalization for adaptively-chosen estimators via stable median," in *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, vol. 65, PMLR, 2017, pp. 728–757 (cit. on pp. 1, 2).

[FS18] V. Feldman and T. Steinke, "Calibrating noise to variance in adaptive data analysis," in *Conference On Learning Theory, COLT 2018*, vol. 75, PMLR, 2018, pp. 535–544 (cit. on pp. 1, 2).

[GGM86] O. Goldreich, S. Goldwasser, and S. Micali, "How to construct random functions," *Journal of the ACM*, vol. 33, no. 4, pp. 792–807, 1986 (cit. on p. 8).

[GL89] O. Goldreich and L. A. Levin, "A hard-core predicate for all one-way functions," in *Proceedings of the Twenty-First Annual ACM Symposium on Theory of Computing*, ser. STOC '89, 1989, 25–32 (cit. on p. 23).

[GMW87] O. Goldreich, S. Micali, and A. Wigderson, "How to play any mental game or A completeness theorem for protocols with honest majority," in *Proceedings of the 19th Annual ACM Symposium on Theory of Computing, STOC 1987*, 1987, pp. 218–229 (cit. on p. 8).

[HILL99] J. Håstad, R. Impagliazzo, L. A. Levin, and M. Luby, "A pseudorandom generator from any one-way function," *SIAM Journal on Computing*, vol. 28, no. 4, pp. 1364–1396, 1999 (cit. on p. 8).

[HMST22] I. Haitner, N. Mazor, J. Silbak, and E. Tsfadia, "On the complexity of two-party differential privacy," in *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, 2022*, ACM, 2022, pp. 1392–1405 (cit. on p. 10).

[HNO+09] I. Haitner, M. Nguyen, S. J. Ong, O. Reingold, and S. Vadhan, "Statistically hiding commitments and statistical zero-knowledge arguments from any one-way function," *SIAM Journal on Computing*, vol. 39, no. 3, pp. 1153–1218, 2009 (cit. on p. 8).

[Hol06] T. Holenstein, "Strengthening key agreement using hard-core sets," Ph.D. dissertation, ETH ZURICH, 2006 (cit. on p. 23).

[HR10] M. Hardt and G. Rothblum, "A multiplicative weights mechanism for privacy-preserving data analysis," in *Proc. 51st Foundations of Computer Science (FOCS)*, IEEE, 2010, pp. 61–70 (cit. on p. 3).

[HU14] M. Hardt and J. Ullman, "Preventing false discovery in interactive data analysis is hard," in *FOCS*, 2014, pp. 454–463 (cit. on pp. 1–5, 8, 12, 17, 18).

[IL89] R. Impagliazzo and M. Luby, "One-way functions are essential for complexity based cryptography (extended abstract)," in *30th Annual Symposium on Foundations of Computer Science, FOCS 1989*, 1989, pp. 230–235 (cit. on p. 7).

[Imp95] R. Impagliazzo, "A personal view of average-case complexity," in *Proceedings of the Tenth Annual Structure in Complexity Theory Conference*, IEEE Computer Society, 1995, pp. 134–147 (cit. on p. 8).

[IR89] R. Impagliazzo and S. Rudich, "Limits on the provable consequences of one-way permutations," in *Annual ACM Symposium on Theory of Computing (STOC)*, 1989, pp. 44–61 (cit. on p. 8).

[JLN+20] C. Jung, K. Ligett, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and M. Shenfeld, "A new analysis of differential privacy's generalization guarantees," in *ITCS*, ser. LIPIcs, vol. 151, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2020, 31:1–31:17 (cit. on p. 1).

[KMNS21] H. Kaplan, Y. Mansour, K. Nissim, and U. Stemmer, "Separating adaptive streaming from oblivious streaming using the bounded storage model," in *CRYPTO (3)*, ser. Lecture Notes in Computer Science, vol. 12827, Springer, 2021, pp. 94–121 (cit. on p. 8).

[KSS22] A. Kontorovich, M. Sadigurschi, and U. Stemmer, "Adaptive data analysis with correlated observations," in *ICML*, ser. Proceedings of Machine Learning Research, vol. 162, PMLR, 2022, pp. 11 483–11 498 (cit. on p. 1).

[LR88] M. Luby and C. Rackoff, "How to construct pseudorandom permutations from pseudorandom functions," *SIAM Journal on Computing*, vol. 17, no. 2, pp. 373–386, 1988 (cit. on p. 8).

[Nao91] M. Naor, "Bit commitment using pseudorandomness," *Journal of Cryptology*, vol. 4, no. 2, pp. 151–158, 1991 (cit. on p. 8).

[NSS+18] K. Nissim, A. D. Smith, T. Steinke, U. Stemmer, and J. R. Ullman, "The limits of post-selection generalization," in *NeurIPS*, 2018, pp. 6402–6411 (cit. on pp. 1, 8).

[Rom90] J. Rompel, "One-way functions are necessary and sufficient for secure signatures," in *Annual ACM Symposium on Theory of Computing (STOC)*, 1990, pp. 387–394 (cit. on p. 8).

[RRST16] R. M. Rogers, A. Roth, A. D. Smith, and O. Thakkar, "Max-information, differential privacy, and post-selection hypothesis testing," in *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016*, 2016, pp. 487–494 (cit. on p. 1).

[RZ16] D. Russo and J. Zou, "Controlling bias in adaptive data analysis using information theory," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016*, vol. 51, JMLR.org, 2016, pp. 1232–1240 (cit. on p. 1).

[Sha84] A. Shamir, "Identity-based cryptosystems and signature schemes," in *Advances in Cryptology, Proceedings of CRYPTO 1984*, vol. 196, 1984, pp. 47–53 (cit. on pp. 5, 10, 11).

[SL19] M. Shenfeld and K. Ligett, "A necessary and sufficient stability notion for adaptive generalization," in *NeurIPS*, 2019, pp. 11 481–11 490 (cit. on p. 1).

[Smi17] A. D. Smith, "Information, privacy and stability in adaptive data analysis," *CoRR*, vol. abs/1706.00820, 2017 (cit. on p. 1).

[SU15a] T. Steinke and J. Ullman, "Between pure and approximate differential privacy," *CoRR*, vol. abs/1501.06095, 2015 (cit. on pp. 1, 2).

[SU15b] T. Steinke and J. Ullman, "Interactive fingerprinting codes and the hardness of preventing false discovery," in *COLT*, 2015, pp. 1588–1628 (cit. on pp. 1–5, 8, 12, 13, 17, 18).

A Proving Theorem 2.4

Theorem 2.4 is an immediate corollary of the following statements.

Theorem A.1 (Key agreement amplification, a corollary of Theorem 7.5 in [Hol06])**.** *If there exists an* $(1 - n^{-\Omega(1)}, 1 - \Omega(1))$*-key agreement protocol, then there exists a fully-secure key-agreement protocol.*

Lemma A.2 (From approximate agreement to a weak key-agreement)**.** *Let* $\alpha, \beta \colon \mathbb{N} \to [0, 1]$ *be efficiently computable functions such that* $\alpha(n)/\beta(n) \leq n^{-\Omega(1)}$ *and* $\alpha(n) \cdot \beta(n) \geq 2^{-n}$ *for large enough* n*. If there exists an* (α, β)*-approximate-agreement protocol, then there exists an* $(1 - n^{-\Omega(1)}, 1 - \Omega(1))$*-key-agreement protocol.*

A.1 Proving Lemma A.2

In the following, for two binary vectors $x, y \in \{0, 1\}^m$, we let $\langle x, y \rangle = \sum_{i=1}^{m} x_i y_i$ (the inner product of x and y), and let $x \oplus y = (x_1 \oplus y_1, \ldots, x_m \oplus y_m)$ (i.e., the bit-wise XOR of x and y). To prove Lemma A.2, we use the following weak version of Goldreich Levin [GL89].

Lemma A.3. *There exists a* PPT *oracle-aided algorithm* Dec *such that the following holds for every* $n \in \mathbb{N}$*. Let* $m \leq n$*,* $x \in \{0,1\}^m$ *and* A *be an algorithm such that*

$$\Pr_{r \sim \{0,1\}^m}[\mathsf{A}(r) = \langle x, r \rangle \bmod 2] > 3/4 + 0.01.$$

Then $\Pr\left[\mathsf{Dec}^{\mathsf{A}}(1^n, 1^m) = x\right] \geq 1 - \mathrm{neg}(n)$*.*

Proof. We use A to decode each bit of x separately. For every i, let $e_i \in \{0,1\}^m$ be the vector that has 1 in the i^{th} entry and 0 everywhere else. Observe that

$$\begin{aligned}
&\Pr_{r \leftarrow \{0,1\}^m}[\{\mathsf{A}(r) = \langle x, r \rangle \bmod 2\} \wedge \{\mathsf{A}(r \oplus e_i) = \langle x, r \oplus e_i \rangle \bmod 2\}] \\
&\geq 1 - 2 \cdot \Pr_{r \leftarrow \{0,1\}^m}[\{\mathsf{A}(r) \neq \langle x, r \rangle \bmod 2\}] \\
&\geq 1/2 + 0.01
\end{aligned}$$

By the linearity of the inner product we deduce that,

$$\Pr_{r \leftarrow \{0,1\}^m}[\mathsf{A}(r) \oplus \mathsf{A}(r \oplus e_i) = x_i] \geq 1/2 + 0.01.$$

Let Dec be the algorithm that for every i, computes $\mathsf{A}(r) \oplus \mathsf{A}(r \oplus e_i)$ for n random values of $r \leftarrow \{0,1\}^m$, and let x'_i be the majority of the outputs. Then Dec outputs $x' = (x'_1, \ldots, x'_n)$. By Hoeffding's inequality Fact 2.1, each x'_i is equal to x_i with all but $e^{-\Omega(n)}$ probability. Since $m \leq n$ we conclude by the union bound that the above is true for all i's simultaneously with all but $\mathrm{neg}(n)$ probability, as required.

□

We now ready to prove Lemma A.2 that transforms an approximate-agreement protocol into a weak key-agreement protocol.

Protocol A.4 (Weak key-agreement protocol $(\mathsf{P}_1, \mathsf{P}_2)$)**.**

Input: 1^n*.*

Access: *An* (α, β)*-approximate-agreement protocol* $(\mathsf{P}'_1, \mathsf{P}'_2)$*.*

Operation:

- *Let* $\gamma = \sqrt{\alpha(n)\beta(n)}$*, and let* $\mathcal{B} = \{-1, -1+\gamma, -1+2\gamma, \ldots, -1+\lfloor 2/\gamma \rfloor \cdot \gamma\}$ *be a division of* $[-1,1]$ *into buckets, each can be represented using* $m = \lceil \log_2(2/\gamma) \rceil$ *bits.*
- *The parties (jointly) emulate* $(\mathsf{P}'_1, \mathsf{P}'_2)$ *on input* 1^n*, where each* P_i *takes the role of* P'_i*. Let* o'_i *be the output of the emulated* P'_i*.*
- P_1 *chooses* $v \leftarrow [0, \gamma]$ *and* $r \leftarrow \{0,1\}^m$ *and sends them to* P_2*.*
- *Each* P_i *computes* $s_i \in \{0,1\}^m$ *as the binary representation of the bucket* $b_i = \mathrm{argmin}_{b \in \mathcal{B}}\{|b - (o'_i + v)|\}$*, and (locally) outputs* $o_i = \langle s_i, r \rangle \bmod 2$*.*

Proof of Lemma A.2. Let $(\mathsf{P}'_1, \mathsf{P}'_2)$ be an (α, β)-approximate-agreement protocol where $\alpha(n)\cdot\beta(n) \geq 2^{-n}$ and $\alpha(n)/\beta(n) \leq n^{-c}$ for a constant $c > 0$ and large enough n. We prove the lemma by showing that Protocol A.4 $(\mathsf{P}_1, \mathsf{P}_2)$ is an $(1 - n^{-c/2}, 0.9)$-key-agreement protocol.

Fix large enough $n \in \mathbb{N}$ and consider a random execution of $(\mathsf{P}_1, \mathsf{P}_2)(1^n)$. Let O'_1, O'_2, V, R, B_1, B_2, O_1, O_2 be the (r.v.'s of the) values of o'_1, o'_2, v, r, b_1, b_2, o_1, o_2 in the execution, let T be the transcript of the execution, and let T' be the transcript of the emulated execution $(\mathsf{P}'_1, \mathsf{P}'_2)(1^n)$ in Step A.4. By the approximate agreement property of $(\mathsf{P}'_1, \mathsf{P}'_2)$, it holds that

$$\Pr\left[\left|O'_1 - O'_2\right| \leq \alpha\right] \geq 1 - \mathrm{neg}(n) \tag{3}$$

Since V is independent of O'_1 and O'_2, it holds that

$$\Pr\left[B_1 = B_2 \mid \left|O'_1 - O'_2\right| \leq \alpha\right] \geq 1 - \alpha/\gamma \geq 1 - n^{-c/2} \tag{4}$$

Hence, the agreement of $(\mathsf{P}_1, \mathsf{P}_2)$ holds by the following computation.

$$\begin{aligned}
\Pr[O_1 = O_2] &\geq \Pr[B_1 = B_2] \\
&\geq \Pr\left[B_1 = B_2 \mid \left|O'_1 - O'_2\right| \leq \alpha\right] \cdot \Pr\left[\left|O'_1 - O'_2\right| \leq \alpha\right] \\
&\geq (1 - n^{-c/2}) \cdot (1 - \mathrm{neg}(n)) \\
&= 1 - n^{-c/2} - \mathrm{neg}(n)
\end{aligned}$$

In order to prove the secrecy of $(\mathsf{P}_1, \mathsf{P}_2)$, assume towards a contradiction that there exists a PPT A such that

$$\Pr[\mathsf{A}(T) = O_1] \geq 0.9 \tag{5}$$

By Lemma A.3, there exists a PPT oracle-aided algorithm Dec such that

$$\Pr\left[\mathsf{Dec}^{\mathsf{A}}(T) = S_1\right] \geq 1 - \mathrm{neg}(n). \tag{6}$$

Let A' be the PPT algorithm that given a transcript t' of the execution of $(\mathsf{P}'_1, \mathsf{P}'_2)$, samples $v \leftarrow [0, \gamma]$ and $r \leftarrow \{0,1\}^m$ (as in Step A.4 of Protocol A.4), computes $t = (v, r, t')$ and outputs the bucket $b \in \mathcal{B}$ that is represented by the binary string $\mathsf{Dec}^{\mathsf{A}}(t) \in \{0,1\}^m$. Since the transcript t induced by $\mathsf{A}'(T')$ is distributed the same as T, we conclude that

$$\begin{aligned}
\Pr\left[\left|\mathsf{A}'(T') - O'_1\right| \leq \beta\right] &\geq \Pr\left[\left|\mathsf{A}'(T') - O'_1\right| \leq 2\gamma\right] \\
&\geq \Pr\left[\left|\mathsf{A}'(T') - O'_1\right| \leq 2\gamma \mid \left|O'_1 - O'_2\right| \leq \alpha\right] - \mathrm{neg}(n) \\
&\geq \Pr\left[\mathsf{Dec}^{\mathsf{A}}(T) = S_1 \mid \left|O'_1 - O'_2\right| \leq \alpha\right] - \mathrm{neg}(n) \\
&\geq \Pr\left[\mathsf{Dec}^{\mathsf{A}}(T) = S_1\right] - \mathrm{neg}(n) \\
&\geq 1 - \mathrm{neg}(n),
\end{aligned}$$

in contradiction to the secrecy property of $(\mathsf{P}'_1, \mathsf{P}'_2)$.

□