# Efficient Adaptive Data Analysis over Dense Data Distributions

**Joon Suk Huh**
Department of Computer Science
University of Wisconsin–Madison, USA
joon@cs.wisc.edu

## Abstract

Modern data workflows are inherently adaptive, repeatedly querying the same dataset to refine and validate sequential decisions, but such adaptivity can lead to overfitting and invalid statistical inference. Adaptive Data Analysis (ADA) mechanisms address this challenge; however, there is a fundamental tension between computational efficiency and sample complexity. For $T$ rounds of adaptive analysis, computationally efficient algorithms typically incur suboptimal $\mathcal{O}(\sqrt{T})$ sample complexity, whereas for arbitrary distributions, algorithms with $\mathcal{O}(\log T)$ sample complexity are computationally intractable under standard cryptographic assumptions. In this work, we shed light on this trade-off by identifying a natural class of data distributions under which both computational efficiency and logarithmic sample complexity in $T$ are achievable. We propose a computationally efficient ADA mechanism that attains $\mathcal{O}(\log T)$ sample complexity, optimal in its dependence on $T$, when the data distribution is dense with respect to a known prior. This setting includes, in particular, feature–label data distributions arising in distribution-specific learning. As a consequence, our mechanism also yields a sample-efficient (i.e., $\mathcal{O}(\log T)$ samples) statistical query oracle in the distribution-specific setting. Moreover, although our algorithm is not based on differential privacy, it satisfies a relaxed privacy notion known as Predicate Singling Out (PSO) security [CN20] whenever the number of its possible update transcripts is polynomial in the sample size. Our results thus reveal an inherent connection between adaptive data analysis and privacy beyond differential privacy.

## 1 Introduction

The inherent adaptivity prevalent in modern data science workflows, while undeniably powerful for discovery and iterative refinement, introduces significant challenges to statistical validity. The repeated interaction with and reuse of the same dataset for both hypothesis generation and subsequent validation can lead to *overfitting* or the generation of *false discoveries* (often colloquially termed "p-hacking" or attributed to "researcher degrees of freedom"). Such issues result in findings that may appear statistically significant within the analyzed sample but fail to generalize reliably to the underlying population from which the data were drawn. This breakdown of classical statistical guarantees contributes directly to the reproducibility crisis observed across empirical scientific fields [Ioa05, GL14]. The core problem, often termed "double-dipping", arises when the same data is used first to inform the analysis strategy and then again to validate the resulting conclusions, thereby coupling the randomness in the inferential target with the randomness in the data used for inference.

Adaptive data analysis (ADA) [DFH$^+$15c] is a modern algorithmic framework aimed at enabling statistically valid inference in adaptive settings, using as few samples as possible. At its core is the design of an intermediary, called an ADA mechanism, that mediates between the data and the analyst, providing answers to adaptively chosen queries while ensuring statistical validity. Since

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

the foundational work of [HU14, DFH+15c], there has been extensive progress in developing algorithmic techniques for preserving validity in adaptive settings [DFH+15a, DFH+15b, CLN+16, RZ16, RRST16, BNS+16, Smi17, FS17, FS18, DSU19, ZH19, JLN+20, FRR20, RRS+20, KSS22, DK22, MR25, Bla25], primarily based on Differential Privacy (DP) [DR14] and/or algorithmic stability [BNS+16, Smi17, SL19, SL23].

In the standard setting, an analyst asks linear queries $q(x) \in \mathbb{R}$ and seeks to estimate expectations of the form $\mathbb{E}_{x\sim\mathcal{D}}[q(x)]$. When the queries are fully adaptive, existing computationally efficient approaches achieve $\mathcal{O}(\sqrt{T})$ sample complexity, where $T$ is the number of query rounds. This rate is tight for *general* data distributions: an $\Omega(\sqrt{T})$ lower bound holds for any *computationally efficient* ADA mechanism under standard cryptographic assumptions, such as the existence of one-way functions [SU15, SU16, BSU17, USN+18, NST23, DSWZ23].

In contrast, data sketching-based methods such as Private Multiplicative Weights (PMW), which maintain an internal sketch of the dataset [HR10], can achieve $\mathcal{O}(\log T)$ sample complexity, but are computationally intractable, as maintaining such a sketch often requires time exponential in the data dimension $n$.

Hence, there is a general tension between computational efficiency and sample efficiency in adaptive data analysis. However, existing hardness results that rule out efficient $o(\sqrt{T})$ sample complexity assume that the mechanism must work for *arbitrary* data distributions. This observation raises a natural question: *Is there a natural class of data distributions $\mathcal{D}$ that admits both $\mathcal{O}(\log T)$ sample complexity and computational efficiency?*

We answer this question in the affirmative. Specifically, we show that for any sufficiently well-spread ("dense") distribution with respect to a known prior (which may be uniform), there exists an ADA mechanism that is both computationally efficient and sample efficient, with sample complexity growing only logarithmically in the number of queries. Such distributions arise naturally in distribution-specific learning [Kha93, BL13, Sha18, BLST25], including distributions over pairs $(x, f(x))$ where $x$ is drawn from a known marginal distribution over features and $f$ is an unknown Boolean function.

### 1.1 Main Results

**Efficient ADA Mechanism for Dense Distributions.** We first present a computationally efficient ADA mechanism for dense data distributions. Density is defined with respect to a known reference (prior) distribution. Specifically, we assume that both the mechanism and the analyst are given a prior distribution $\mathcal{D}_g$ generated by a known, efficiently computable function $g : \{0,1\}^n \to \mathcal{I}$ over a finite domain $\mathcal{I}$. We then consider an *unknown* target distribution $\mathcal{D}$ that is $\lambda$-dense in $\mathcal{D}_g$, meaning that $\Pr_{I\sim\mathcal{D}}[I = i] \le \frac{1}{\lambda}\Pr_{I\sim\mathcal{D}_g}[I = i]$ for all $i \in \mathcal{I}$.

Intuitively, this condition implies that one could sample from $\mathcal{D}$ using $\mathcal{D}_g$ via rejection sampling with acceptance probability at least $\lambda$. However, this form of "rejection samplability" does not imply that the analyst can perform data analysis over the unknown distribution $\mathcal{D}$, since the required acceptance rule depends on the unknown target. Such an unknown rule arises, for example, when a filter retains records of a known population with average acceptance probability at least $\lambda$; the retained records then follow a $\lambda$-dense distribution.

**Theorem 1.1** (Restatement of Theorem 3.1 and Corollary 3.3)**.** *Let $g : \{0,1\}^n \to \mathcal{I}$ be an efficiently computable function. There exists an ADA mechanism, depending only on $g$, $\lambda$, $\epsilon$, $\delta$, and $T$, with the following guarantee. For any distribution $\mathcal{D}$ over $\mathcal{I}$ that is $\lambda$-dense with respect to the distribution $\mathcal{D}_g$ generated by $g$, the mechanism answers $\mathbb{E}_{k\sim\mathcal{D}}[q_t(k)]$ to within $\epsilon$ accuracy for any sequence of $T$ adaptively chosen queries $(q_t)_{t\in[T]}$, where each $q_t : \mathcal{I} \to \{-1, 1\}$, with probability at least $1-\delta$, using $M \in \mathcal{O}(\lambda^{-1}\epsilon^{-4}\log(T/\epsilon) + \epsilon^{-2}\log(1/\delta))$ i.i.d. samples from $\mathcal{D}$. For queries evaluable in $\mathsf{poly}(n)$ time, the runtime is $\mathsf{poly}(n, \epsilon^{-1}, \lambda^{-1}, T, \log(1/\delta))$.*

Notably, the sample complexity is independent of the ambient data dimension $n$, and its logarithmic dependence on $T$ is optimal even for non-adaptive queries at constant density (Proposition 3.4). More generally, if some surrogate $\mathcal{D}'$ matches the analyst's queries within $\epsilon/16$ and its likelihood ratio $w = d\mathcal{D}'/d\mathcal{D}_g$ satisfies $\mathbb{E}_{\mathcal{D}_g} w^2 \le \rho^2$, the sample complexity is $\mathcal{O}(\rho^2\epsilon^{-4}\log(T/\epsilon) + \epsilon^{-2}\log(1/\delta))$; under pointwise $\lambda$-density, $\mathbb{E}_{\mathcal{D}_g} w^2 \le 1/\lambda$, so $\rho = \lambda^{-1/2}$ suffices.

**Sample-efficient Distribution-Specific Statistical Query Oracle.** Why, then, do we care about dense target distributions? A natural example arises in distribution-specific learning. Let $X \sim \mathcal{D}_h$ for a known, efficiently samplable feature marginal and $Y = f(X) \in \{-1, 1\}$ for an unknown labeling function. This labeled distribution is $1/2$-dense with respect to the prior that pairs $X$ with an independent uniform label. A known feature sampler of this kind arises, for example, in experimental designs and simulations where features come from a known process and outcomes are unknown.

Our mechanism yields a distribution-specific Statistical Query (SQ) oracle [Kea98] using one labeled sample for all adaptive calls (Theorem 3.5). Bounded losses, gradient coordinates, and conditional label statistics reduce to such queries; in §2.2 we give the reduction and compare the resulting oracle with classical SQ simulation.

**Privacy Guarantee.** Many existing ADA mechanisms are built on differential privacy (DP) and thus often inherit formal DP guarantees. In contrast, our mechanism is not based on DP and does not provide DP guarantees. Nevertheless, perhaps surprisingly, it still offers a meaningful privacy guarantee in terms of *Predicate Singling Out* (PSO) security [CN20], a notion that is strictly weaker than DP and was designed to capture the "singling out" risk addressed by the European Union's General Data Protection Regulation (GDPR). Specifically, the success probability of a PSO adversary is at most the number of possible update transcripts times the largest success probability of an adversary that sees no output, so PSO security holds when this number is polynomial in the sample size, for example at constant accuracy and density with $T = \mathsf{poly}(M)$ (Theorem 5.2). This bound follows from an argument that builds on a key lemma from our sample complexity analysis and, notably, requires no explicit noise addition.

### 1.2 Technical Overview

Unlike many computationally efficient ADA mechanisms that achieve $\mathcal{O}(\sqrt{T})$ sample complexity, our mechanism does not rely on differential privacy or algorithmic stability arguments. Instead, it builds on an iterative data sketching strategy inspired by the Private Multiplicative Weights (PMW) algorithm [HR10]. The goal is to maintain an internal sketch of the dataset $S \in \mathcal{I}^M$ such that linear statistical queries evaluated on the sketch closely approximate their values on $S$. While PMW achieves this by maintaining a distribution over the data universe and updating it via multiplicative weights, this approach incurs computational cost proportional to the domain size $2^n$, which is intractable.

We avoid this intractability by using (unconstrained) Online Gradient Descent (OGD) [Zin03] in the seed space $\mathbb{R}^N$, where $N = 2^n$. Through the known prior, we represent a surrogate $\mathcal{D}'$ with likelihood ratio $w = d\mathcal{D}'/d\mathcal{D}_g$ and $\mathbb{E}_{\mathcal{D}_g} w^2 \le \rho^2$ as the vector with coordinates $w(g(r))/(\rho\sqrt{N})$, which lies in the Euclidean unit ball and is the comparator in our analysis. Specifically, our algorithm can be viewed as a lazy version of OGD, where an update is performed only when the approximation error exceeds a prescribed threshold. This lazy update rule, together with the standard regret bound for OGD, implies that the number of such updates, and hence the number of vectors stored in the sketch, is bounded by $\mathcal{O}(\rho^2\epsilon^{-2})$. Crucially, the intractability of PMW is avoided because each stored vector is a scaled query function, so answers can be reconstructed efficiently from the sketch via randomized inner-product estimation at uniformly random seeds.

Equivalently, the current sketch minimizes the sum of the recorded linearized losses plus a fixed quadratic regularizer, so the mechanism can also be viewed as follow-the-regularized-leader (FTRL) [McM17]; Equation (B.4) states the exact identity.

The bounded number of updates (mistakes) of order $\mathcal{O}(\rho^2\epsilon^{-2})$ in constructing the internal sketch of the dataset implies the existence of a succinct transcript of the mechanism's update behavior, of size $\mathcal{O}(\rho^2\epsilon^{-2}\log(T/\epsilon))$ bits. As this transcript can simulate the mechanism's exact behavior, it can be used to reconstruct all queries asked by $\mathcal{A}$, using only the internal randomness of the mechanism and the analyst, and without access to the sample $S$. This transcript plays a central role in ensuring statistical validity. Since it determines the entire query sequence, we can apply standard concentration arguments, namely Hoeffding's inequality and the union bound, over the small space of possible transcripts. This yields uniform convergence, guaranteeing that $|q_t(S) - q_t(\mathcal{D})| \in \mathcal{O}(\epsilon)$ for all $t \in [T]$ with high probability, with sample size $|S| \in \mathcal{O}(\rho^2\epsilon^{-4}\log(T/\epsilon) + \epsilon^{-2}\log(1/\delta))$.

**Game 1:** $\mathsf{AQ}_{N,M,T}[\mathcal{A} \rightleftharpoons \mathcal{M}]$: Adaptive Query Game

$\mathcal{A}$ chooses a distribution $\mathcal{D}$ over $\mathcal{I}$ unknown to $\mathcal{M}$.
A multiset $S$ of $M$ i.i.d. samples from $\mathcal{D}$ is given to $\mathcal{M}$ but not to $\mathcal{A}$.
**for** $t \in [T]$ **do**
  $\mathcal{A}$ chooses a query $q_t \in \{-1, 1\}^{\mathcal{I}}$ given $(q_1, a_1, \ldots, q_{t-1}, a_{t-1})$.
  $\mathcal{M}$ outputs an answer $a_t \in [-1, 1]$ given $(q_1, a_1, \ldots, q_{t-1}, a_{t-1}, q_t)$.
**end**

### 1.3 Related Work

Many algorithmic frameworks for adaptive data analysis (ADA) are based on differential privacy or algorithmic stability, as studied in [BNS+16, Smi17, FS18, SL19, JLN+20, Bla25]. These approaches provide strong generalization guarantees under adaptivity and run efficiently, but typically incur a sample complexity of $\mathcal{O}(\sqrt{T})$.

Closer to our approach, the idea of answering queries by reconstructing (or sketching) a dataset was introduced in [BLR13], which used the exponential mechanism [MT07] both to reconstruct the dataset and to privately answer *non-adaptive* queries. This was extended in the seminal work of [HR10], which proposed the Private Multiplicative Weights (PMW) algorithm, based on the multiplicative weights framework [FS99, AHK12], to handle adaptive queries while satisfying differential privacy. Notably, while PMW is designed as a differentially private algorithm, it also serves as an ADA mechanism with $\mathcal{O}(\log T)$ sample complexity and was later extended beyond linear queries [Ull15]. However, it is computationally intractable in general, as it requires sampling from Gibbs distributions over exponentially large domains like $\{0, 1\}^n$.

Hardt and Rothblum [HR10] also give an efficient variant of PMW for pseudo-smooth databases. A $\xi$-smooth histogram puts mass at most $\xi$ on each point, so smoothness is $\lambda$-density relative to the uniform distribution with $\lambda = 1/(\xi N)$; pseudo-smoothness relative to a query class parallels our pseudodensity. The accuracy guarantee of this variant [HR10, Theorem I.2] holds only for query sequences fixed in advance. Their general PMW mechanism allows adaptive queries but runs in time nearly linear in the universe size. Our result provides population accuracy for adaptive queries under a shared efficient prior sampler, in time polynomial in $n$. For distribution-specific SQ algorithms, we compare our implementation with classical SQ simulation [Kea98] in §2.2.

### 1.4 Organization

In §2, we introduce the adaptive data analysis framework and define (pseudo)dense data distributions. In §3, we present our efficient ADA mechanism, denoted $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$. In §4, we establish the correctness and sample complexity guarantees of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$. In §5, we introduce Predicate Singling Out (PSO) and bound the success probability of PSO attacks. We discuss limitations in §6.

## 2 Preliminaries

In this section, we provide the necessary background on adaptive data analysis and introduce key definitions used throughout the paper.

**Notation.** We denote the set $\{1, \ldots, N\}$ by $[N]$. The sign function $\mathsf{sign}(x)$ returns $+1$ if $x \geq 0$ and $-1$ otherwise. The set of non-negative real numbers is denoted by $\mathbb{R}_+$. For any vector $v \in \mathbb{R}^N$, we write $\|v\|$ to denote its Euclidean norm. The $N$-dimensional Euclidean unit ball is defined as $\mathbb{B}^N := \{v \in \mathbb{R}^N : \|v\| \leq 1\}$. For any finite set $S$, $U_S$ denotes the uniform distribution over $S$. Finally, for any function $q : \mathcal{I} \to \mathbb{R}$ and any distribution $\mathcal{D}$ over $\mathcal{I}$, we define $q(\mathcal{D}) := \mathbb{E}_{i \sim \mathcal{D}}[q(i)]$ as the expectation of $q$ under $\mathcal{D}$. For a multiset $S = (I_1, \ldots, I_M)$ of records in $\mathcal{I}$, we write $q(S) := M^{-1} \sum_{j=1}^{M} q(I_j)$, counting repeated records with their multiplicities. For the sake of presentation, we identify the $n$-bit Boolean cube $\{0, 1\}^n$ with $[N]$, where $N = 2^n$, throughout the paper. We take $T \geq 1$, $0 < \epsilon, \delta \leq 1/2$, and $0 < \lambda \leq 1$.

### 2.1 Problem Setting: Adaptive Data Analysis

An adaptive data analysis (ADA) mechanism $\mathcal{M}$ is an algorithm that holds a sufficient amount of i.i.d. data $S = (I_1, \ldots, I_M)$ drawn from an *unknown* distribution $\mathcal{D}$ over a finite domain $\mathcal{I}$, and provides answers $a_t$ to a sequence of adaptively chosen Boolean queries $(q_t)_{t\in[T]}$, where each $q_t : \mathcal{I} \to \{-1, 1\}$, selected by an analyst (or adversary) $\mathcal{A}$. The goal is to ensure that these answers are $\epsilon$-accurate; that is, $|a_t - q_t(\mathcal{D})| \le \epsilon$ for all $t \in [T]$.

In the non-adaptive setting, where all queries are chosen in advance, it suffices for $\mathcal{M}$ to hold $|S| \in \mathcal{O}(\log T)$ samples, by standard concentration arguments (e.g., Hoeffding's inequality and the union bound). Our goal is to retain this $\mathcal{O}(\log T)$ sample complexity even when queries are chosen adaptively, under certain assumptions on the data distribution, while ensuring that the ADA mechanism remains *computationally efficient*; that is, it runs in time $\mathsf{poly}(\epsilon^{-1}, \lambda^{-1}, \log N, |S|, T, \log(1/\delta))$, with $\lambda^{-1}$ replaced by $\rho$ under the bounded-$L_2$ condition below.

We formalize adaptive data analysis as a game between the ADA mechanism $\mathcal{M}$ and the analyst (or adversary) $\mathcal{A}$, referred to as the *Adaptive Query Game*, depicted in Game 1. Note that allowing $\mathcal{A}$ to choose the distribution $\mathcal{D}$ in Game 1 only strengthens $\mathcal{A}$; thus, our result also applies to the more natural setting in which $\mathcal{D}$ is *unknown* to both the analyst $\mathcal{A}$ and the mechanism $\mathcal{M}$.

Let $\Pr_{\mathsf{AQ}_{N,M,T}[\mathcal{A}\rightleftharpoons\mathcal{M}]}$ denote the probability distribution over all random variables (such as the $a_t$s and $q_t$s) that are realized during the execution of the game $\mathsf{AQ}_{N,M,T}[\mathcal{A} \rightleftharpoons \mathcal{M}]$. We then define accuracy of $\mathcal{M}$ as follows:

**Definition 2.1** (Accuracy)**.** A mechanism $\mathcal{M}$ is said to be $(\epsilon, \delta)$-accurate with respect to $\mathcal{A}$ with sample complexity $M_0$ if, for all $M \ge M_0$,

$$\Pr_{\mathsf{AQ}_{N,M,T}[\mathcal{A}\rightleftharpoons\mathcal{M}]}\Big[\forall t \in [T],\ |a_t - q_t(\mathcal{D})| \le \epsilon\Big] \ge 1 - \delta,$$

where the probability is taken over $S \sim \mathcal{D}^M$ and the internal randomness of both $\mathcal{M}$ and $\mathcal{A}$.

**Dense Data Distribution.** We assume that both the analyst and the mechanism have access to a function $g : [N] \to \mathcal{I}$ for some finite domain $\mathcal{I}$, and hence can sample from the distribution $\mathcal{D}_g$, where $I \sim \mathcal{D}_g$ is given by $I := g(X)$ with $X \sim U_{[N]}$. We further assume that $g$ is efficiently computable, namely in $\mathsf{poly}(\log N)$ time. We say that a distribution $\mathcal{D}$ over $\mathcal{I}$ is $(\lambda, \mathcal{D}_g)$-*dense* if $\Pr_{I\sim\mathcal{D}}[I = i] \le \frac{1}{\lambda}\Pr_{I\sim\mathcal{D}_g}[I = i]$ for all $i \in \mathcal{I}$.

**Examples of Dense Distributions.** If a record $z \sim \mathcal{D}_g$ is accepted with an unknown probability $a(z) \in [0, 1]$ and $\mathbb{E}_{\mathcal{D}_g} a \ge \lambda$, the distribution of accepted records satisfies $\mathcal{D}(z)/\mathcal{D}_g(z) = a(z)/\mathbb{E}_{\mathcal{D}_g} a \le 1/\lambda$. This includes conditioning on an event of prior probability at least $\lambda$. Bounded reweighting, $\mathcal{D} = w\mathcal{D}_g$ with $0 \le w \le 1/\lambda$ and $\mathbb{E}_{\mathcal{D}_g} w = 1$, gives the same condition. A component of a known mixture $\mathcal{D}_g = \alpha\mathcal{D} + (1 - \alpha)Q$ is also $\lambda$-dense whenever $\alpha \ge \lambda$. The mechanism needs the shared sampler and a bound on the density parameter; it does not need either probability mass function, the acceptance rule, or a certificate of density.

Our algorithm, in fact, applies to a slightly broader class of distributions that includes not only truly dense distributions but also *pseudodense* ones, namely distributions that are indistinguishable from dense distributions from the perspective of the analyst. We formalize this notion below.

**Definition 2.2** (Pseudodensity)**.** For any adversary $\mathcal{A}$, let $\mathcal{Q}^{\mathcal{A}} \subseteq \{-1, 1\}^{\mathcal{I}}$ denote the set of all queries that $\mathcal{A}$ may output at any point, on any input, and under any internal randomness. We say that a distribution $\mathcal{D}$ over $\mathcal{I}$ is $(\epsilon, \lambda, \mathcal{D}_g)$-*pseudodense* with respect to $\mathcal{A}$ if there exists a $(\lambda, \mathcal{D}_g)$-dense distribution $\mathcal{D}'$ over $\mathcal{I}$ such that $|q(\mathcal{D}) - q(\mathcal{D}')| \le \epsilon$ for all $q \in \mathcal{Q}^{\mathcal{A}}$. Moreover, if $\mathcal{A}$ almost surely selects a distribution $\mathcal{D}$ that is $(\epsilon, \lambda, \mathcal{D}_g)$-pseudodense with respect to itself, we say that $\mathcal{A}$ is an $(\epsilon, \lambda, \mathcal{D}_g)$-*pseudodense challenger*.

Pseudodensity is relative to all queries in $\mathcal{Q}^{\mathcal{A}}$, including those on histories that are not realized. For example, if queries on records $(u, v)$ depend only on $u$, a target with the same $u$-marginal as the prior is $(0, 1, \mathcal{D}_g)$-pseudodense, even if its conditional distribution of $v$ is concentrated. For an unrestricted Boolean query family, the condition instead requires $2\,\mathrm{TV}(\mathcal{D}, \mathcal{D}') \le \epsilon$.

**A Bounded-$L_2$ Formulation.** More generally, the analysis only needs a surrogate $\mathcal{D}' \ll \mathcal{D}_g$ whose likelihood ratio $w = d\mathcal{D}'/d\mathcal{D}_g$ satisfies, for a given bound $\rho \geq 1$,

$$\mathop{\mathbb{E}}_{\mathcal{D}_g}[w^2] \leq \rho^2, \qquad |q(\mathcal{D}) - q(\mathcal{D}')| \leq \epsilon/16 \quad \text{for all } q \in \mathcal{Q}^{\mathcal{A}}. \tag{2.1}$$

Pointwise $\lambda$-density implies $\mathbb{E}_{\mathcal{D}_g}[w^2] \leq \lambda^{-1}\,\mathbb{E}_{\mathcal{D}_g}[w] = \lambda^{-1}$. Thus, $(\epsilon/16, \lambda, \mathcal{D}_g)$-pseudodensity satisfies (2.1) with $\rho = \lambda^{-1/2}$. We use the scale $\kappa = 1/\rho$ in the algorithm. Neither $w$ nor $\mathcal{D}'$ is an input, and $\mathcal{D}$ itself need not be absolutely continuous with respect to $\mathcal{D}_g$.

**Working in the Seed Space.** The prior generator $g$ is fixed throughout the AQ game. For each query $q_t : \mathcal{I} \to \{-1, 1\}$, write $\tilde{q}_t = q_t \circ g \in \{-1, 1\}^N$. The internal sketch uses these vectors in seed space; the sample mean $q_t(S)$ is always evaluated on the observed records in $\mathcal{I}$. This requires evaluating $g$, but never inverting it or recovering seeds for target samples.

### 2.2 Application: Implementing Distribution-Specific Statistical Query Oracle

The above setting also captures a standard distribution-specific learning scenario. Consider a distribution $\mathcal{D}_{h,f}$ over labeled examples $(X, Y)$, where $X$ is drawn from a *known* marginal distribution $\mathcal{D}_h$ over a finite feature space $\mathcal{X}$, generated by an efficient sampler $h$, and $Y \coloneqq f(X) \in \{-1, 1\}$ for an unknown labeling function $f : \mathcal{X} \to \{-1, 1\}$. Define a prior distribution $\mathcal{D}_g$ by letting $I \sim \mathcal{D}_g$ be given as $I \coloneqq g(R) \coloneqq (h(R_{1:n-1}), 2R_n - 1)$ for a uniformly random seed $R \sim U_{[N]}$. Equivalently, $\mathcal{D}_g$ samples a feature vector $X \sim \mathcal{D}_h$ together with an independent uniform label in $\{-1, 1\}$.

Under this construction, the true labeled distribution $\mathcal{D}$ is $(1/2)$-dense with respect to $\mathcal{D}_g$. Consequently, an ADA mechanism designed for $(\lambda, \mathcal{D}_g)$-dense distributions can be interpreted as providing a *distribution-specific* statistical query interface for the labeled distribution $\mathcal{D}_{h,f}$.

**Definition 2.3** (Distribution-specific SQ oracle [Kea98])**.** Fix a known distribution $\mathcal{D}_h$ over a feature space $\mathcal{X}$ and an unknown labeling function $f : \mathcal{X} \to \{-1, 1\}$. Let $\mathcal{D}_{h,f}$ denote the induced distribution over $\mathcal{X} \times \{-1, 1\}$ obtained by sampling $X \sim \mathcal{D}_h$ and setting $Y \coloneqq f(X)$. A distribution-specific statistical query oracle, denoted $\mathsf{DSQ}_h(f, \epsilon)$, takes as input a query function $\phi : \mathcal{X} \times \{-1, 1\} \to [-1, 1]$ and returns a value $v$ satisfying

$$|v - \phi(\mathcal{D}_{h,f})| \leq \epsilon.$$

An algorithm that interacts with the unknown $f$ only through calls to $\mathsf{DSQ}_h(f, \epsilon)$ is called a distribution-specific SQ algorithm (with respect to $\mathcal{D}_h$).

DSQ queries take values in $[-1, 1]$, whereas the queries in Game 1 are Boolean; we handle this mismatch by a standard randomization. For a bounded query $\phi_t$, draw independent thresholds $U_{t,j} \sim \mathrm{Unif}[-1, 1]$ and set $q_{t,j}(x, y) = \mathsf{sign}(\phi_t(x, y) - U_{t,j})$, so that $\mathbb{E}_{U_{t,j}}[q_{t,j}(x, y)] = \phi_t(x, y)$. Each threshold is fixed for all evaluations of its query. Averaging the answers to $B = \Theta(\epsilon^{-2} \log(T/\delta))$ such Boolean queries implements one DSQ call. All $BT$ queries use the same labeled sample; Appendix A gives the error and confidence calculation.

**Motivation for Linear Queries.** The expectation of a bounded loss, or of a bounded coordinate of its gradient, is a linear query after scaling to $[-1, 1]$. Such queries can be issued adaptively by gradient-based optimization methods [FGV21]. Each coordinate counts as a separate call. Conditional label statistics also reduce to linear queries: for bounded $\psi$ and a subgroup $A \subseteq \mathcal{X}$,

$$\mathbb{E}[\psi(Y) \mid X \in A] = \frac{\mathbb{E}[\psi(Y)\mathbf{1}\{X \in A\}]}{\mathcal{D}_h(A)}.$$

The numerator is a linear query and the denominator can be estimated from the known sampler. The subgroup mass $\mathcal{D}_h(A)$ must be bounded away from zero at the desired accuracy scale, since division amplifies estimation error.

**Comparison with SQ Simulation.** Classical SQ simulation already reuses one sample by uniform convergence over the query class [Kea98, Theorem 1]; the bound depends on the class size or its VC dimension. A fresh-sample implementation instead uses $\mathcal{O}(T\epsilon^{-2} \log(T/\delta))$ labeled examples. Using the known marginal, we control the sample complexity through the number $T$ of adaptive calls, with no complexity term for the query class. Our guarantee still requires efficient query evaluation and sampling, and its dependence on the accuracy is worse than that of uniform convergence.

**Algorithm 1:** $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$: Adaptive Query Answering via Online Gradient Descent

1 **Oracle:** Adversary $\mathcal{A}$.
2 **Input:** Data $S \in \mathcal{I}^M$, prior generator $g$, accuracy $\epsilon$, failure probability $\delta$, and norm bound $\rho \geq 1$; set $\kappa = 1/\rho$.
3 $\mathcal{G} \leftarrow \emptyset$ ▷ $\mathcal{G}$ is a maintained multiset of gradients.
4 **for** $t \in [T]$ **do**
5 Get a query $q_t$ from $\mathcal{A}$ and set $\tilde{q}_t = q_t \circ g$.
6 $\hat{u}_t \leftarrow \mathsf{IP}_{\mathsf{est}}(\tilde{q}_t, \mathcal{G}; \kappa\epsilon/8, \delta/2T)$ ▷ $\mathsf{IP}_{\mathsf{est}}$ is given in Alg. 2
7 **if** $|\hat{u}_t - \kappa q_t(S)| > \kappa\epsilon/2$ **and** $|\mathcal{G}| \leq 128(\kappa\epsilon)^{-2}$ **then**
8 $s_t \leftarrow \mathsf{sign}(\hat{u}_t - \kappa q_t(S))$
9 $\mathcal{G} \leftarrow \mathcal{G} \cup \{-\kappa\epsilon s_t \tilde{q}_t/16\}$ ▷ Update $\mathcal{G}$.
10 $a_t \leftarrow$ Round $q_t(S)$ to the nearest multiple of $\epsilon/2$.
11 **end**
12 **else**
13 $a_t \leftarrow \hat{u}_t/\kappa$
14 **end**
15 Set $a_t \leftarrow \min\{1, \max\{-1, a_t\}\}$ and send $a_t$ to $\mathcal{A}$.
16 **end**

**Algorithm 2:** $\mathsf{IP}_{\mathsf{est}}(q, \mathcal{G}; \epsilon, \delta)$

**Input:** Query $q \in \{-1, 1\}^N$, gradient vectors $\mathcal{G}$, accuracy $\epsilon > 0$ and failure probability $\delta > 0$.
**Output:** An $\epsilon$-accurate estimate of $\langle q, y \rangle / N$, where $y \coloneqq \sum_{v \in \mathcal{G}} v$.
Sample $(i_1, \ldots, i_L) \overset{iid}{\sim} U_{[N]}$ where $L \coloneqq \lceil 2\epsilon^{-2} T^2 \ln \frac{2T}{\delta} \rceil$.
**return** $\frac{1}{L} \sum_{v \in \mathcal{G}} \sum_{k=1}^{L} q(i_k) v(i_k)$

## 3 ADA Mechanism via Online Gradient Descent

In this section, we present our main contribution, the ADA mechanism $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ (Alg. 1).

**Overview of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$.** Given samples $S \in \mathcal{I}^M$, $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ aims to answer queries using an approximation (sketch) of $S$ whenever possible, rather than answering with $q_t(S)$ directly for every $t$. To enable this, $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ maintains a multiset $\mathcal{G} \subseteq \mathbb{R}^N$ consisting of vectors proportional to previous queries in seed space, $\tilde{q}_t \in \{-1, 1\}^N$, interpreted as vectors in $\mathbb{R}^N$. Internally, each of these vectors is proportional to a subgradient of the convex objective function $\hat{f}_t(x) \coloneqq \left| \left\langle \tilde{q}_t/\sqrt{N}, x \right\rangle - \kappa q_t(S) \right|$. For this reason, we refer to the set $\mathcal{G}$ as the collection of "gradient" vectors.

In each round, at Line 6, $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ attempts to approximate the desired answer using the $\mathsf{IP}_{\mathsf{est}}$ subroutine, defined in Alg. 2, based on the current collection of gradients $\mathcal{G}$. The desired answer is encoded in $\langle \tilde{q}_t, y_t \rangle / N$ for $y_t \coloneqq \sum_{v \in \mathcal{G}} v$, which is designed to approximate $\kappa q_t(S)$; we call $x_t \coloneqq y_t/\sqrt{N}$ the normalized sketch. The subroutine $\mathsf{IP}_{\mathsf{est}}$ approximates $\langle \tilde{q}_t, y_t \rangle / N$ by taking the sample mean of $\tilde{q}_t(i) y_t(i)$. This randomized estimation avoids the prohibitive cost of full-dimensional computation and ensures that the running time depends on $N$ only through $\mathsf{poly}(\log N)$.

After estimating the tentative answer $\hat{u}_t/\kappa$ in Line 6, $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ checks whether this estimate is sufficiently close to $q_t(S)$. If so, it outputs $a_t = \hat{u}_t/\kappa$; otherwise, it updates $\mathcal{G}$ according to Line 9, effectively performing a gradient step on $\hat{f}_t$, and returns the answer using $q_t(S)$. We refer to this mechanism as AQA-OGD, as it follows an online gradient descent procedure to maintain an approximate representation of $S$.

The key insight lies in the second condition of Line 7. Although $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ evaluates $q_t(S)$ in every round, it updates $\mathcal{G}$ at most $\mathcal{O}(\rho^2\epsilon^{-2})$ times, independent of the total number of queries $T$ or the domain size $N$. As shown in §4, this limitation is crucial to our analysis, ensuring that the accuracy of each $a_t$ is preserved via uniform concentration of $q_t(S) \approx q_t(\mathcal{D})$ for all $t \in [T]$.

**Accuracy and Efficiency of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$.** We now provide the computational efficiency and sample complexity guarantees for $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$. We begin with its computational efficiency. The construction of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ in Algorithms 1 and 2 immediately yields the following bound.

**Theorem 3.1. [Computational Efficiency]** *Under Game 1, given that $\mathcal{A}$ issues queries computable in $\mathsf{poly}(\log N)$ time, $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ runs in time $\mathsf{poly}(\epsilon^{-1}, \rho, \log N, M, T, \log(1/\delta))$.*

More importantly, we state the accuracy and sample complexity guarantees for $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ in the following theorem, which is sketched in §4 and proved in Appendix B.1.

**Theorem 3.2. [Sample Efficiency]** *If $\mathcal{A}$ almost surely chooses a distribution $\mathcal{D}$ for which some surrogate $\mathcal{D}'$ satisfies* (2.1) *with respect to its query family $\mathcal{Q}^{\mathcal{A}}$, then $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$, run with input $\rho$, is $(\epsilon, \delta)$-accurate with respect to $\mathcal{A}$, with any sample size $M \geq C\big(\rho^2\epsilon^{-4}\log(T/\epsilon) + \epsilon^{-2}\log(1/\delta)\big)$, where $C$ is a universal constant.*

**Corollary 3.3. [Dense and Pseudodense Distributions]** *If $\mathcal{A}$ is an $(\epsilon/16, \lambda, \mathcal{D}_g)$-pseudodense challenger, $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ with $\rho = \lambda^{-1/2}$ is $(\epsilon, \delta)$-accurate for $M \geq C\big(\lambda^{-1}\epsilon^{-4}\log(T/\epsilon) + \epsilon^{-2}\log(1/\delta)\big)$.*

The corollary follows from $\mathbb{E}_{\mathcal{D}_g}[w^2] \leq 1/\lambda$. The logarithmic dependence on $T$ is necessary even for non-adaptive queries (Appendix D).

**Proposition 3.4. [Lower Bound]** *For $T \geq 2$ and $0 < \epsilon \leq 1/8$, there is a family of $2/3$-dense distributions relative to the uniform prior on $\{-1,1\}^T$ such that answering the $T$ fixed coordinate queries simultaneously within $\epsilon$, with probability at least $2/3$, requires $M = \Omega(\epsilon^{-2}\log T)$ samples.*

**Sample-efficient Implementation of Distribution-specific SQ Oracle.** Our ADA mechanism also yields a sample-efficient implementation of a distribution-specific SQ oracle. As discussed in §2.2, $\mathcal{D}_{h,f}$ is $1/2$-dense relative to the prior with independent uniform labels. Applying Corollary 3.3 to the Booleanized queries gives an implementation that reuses one labeled sample (Appendix A).

**Theorem 3.5. [Sample-efficient DSQ Oracle from $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$]** *Fix a feature distribution $\mathcal{D}_h$ over a finite domain $\mathcal{X}$ generated by an efficient sampler $h$, and an unknown labeling function $f : \mathcal{X} \to \{-1, 1\}$. Let $\mathcal{D}_{h,f}$ denote the induced distribution over $(X, Y)$ given by $X \sim \mathcal{D}_h$ and $Y = f(X)$. Then there exists an implementation of the distribution-specific SQ oracle $\mathsf{DSQ}_h(f, \epsilon)$ with the following guarantee: for any analyst issuing a sequence of $T$ adaptively chosen queries $\phi_1, \ldots, \phi_T$ with $\phi_t : \mathcal{X} \times \{-1, 1\} \to [-1, 1]$, the implementation answers all queries simultaneously to within $\epsilon$ accuracy with probability at least $1 - \delta$, using*

$$M \in \mathcal{O}\big(\epsilon^{-4}\log(BT/\epsilon) + \epsilon^{-2}\log(1/\delta)\big), \quad B = \Theta(\epsilon^{-2}\log(T/\delta))$$

*i.i.d. samples from $\mathcal{D}_{h,f}$, reused for all calls. If $h$ and each $\phi_t$ can be evaluated in $\mathsf{poly}(\log N)$ time, the runtime is $\mathsf{poly}(\log N, \epsilon^{-1}, T, \log(1/\delta))$.*

## 4 Sample Complexity Analysis

In this section, we analyze the sample complexity of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ as stated in Theorem 3.2.

Our analysis has two parts. First, the bounded update transcript gives uniform concentration of $q_t(S)$ around $q_t(\mathcal{D})$, even though the queries are adaptive. Second, the regret bound for OGD shows that the update budget is sufficient, so every answer remains close to $q_t(S)$.

### 4.1 Proof Sketch of Theorem 3.2

Here we provide a proof sketch of our main result, Theorem 3.2, which establishes the sample complexity of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$. The full proof appears in Appendix B.1.

**Step 1. Reconstructing queries from succinct transcript.** A central component of our analysis is the procedure $\mathsf{Replay}\big(Z;\mathcal{A},\omega\big)$, described in Alg. 3 in Appendix B, which "replays" all queries issued by $\mathcal{A}$ during the execution of $\mathsf{AQ}_{N,M,T}[\mathcal{A} \rightleftharpoons \mathcal{M}_{\mathsf{AQA\text{-}OGD}}]$, when both $\mathcal{A}$ and $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ are run with *fixed* internal randomness $\omega_{\mathcal{A}}$ and $\omega_{\mathcal{M}}$, respectively. The main input to Replay is a set of key-value pairs $Z \coloneqq \{(t_1,(a_{t_1},s_{t_1})),\dots\}$, where each key $t_i$ denotes a round in which $\mathcal{G}$ was updated, and the corresponding value $(a_{t_i},s_{t_i})$ is the answer given to $\mathcal{A}$ and the sign of the update, $s_{t_i} \coloneqq \mathsf{sign}(\hat{u}_{t_i} - \kappa q_{t_i}(S))$, as recorded in Line 8 of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ (Alg. 1) at round $t_i$. We refer to $Z$ as the update transcript.

Given fixed randomness $\omega \coloneqq (\omega_{\mathcal{A}},\omega_{\mathcal{M}})$, samples $S$ and the analyst's algorithm $\mathcal{A}$, by construction, there exists a transcript $Z(\omega,S)$ such that $\mathsf{Replay}(Z(\omega,S);\mathcal{A},\omega)$ faithfully reproduces all queries issued by $\mathcal{A}$, assuming $\mathcal{A}$ and $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ are executed with frozen randomness $\omega_{\mathcal{A}}$ and $\omega_{\mathcal{M}}$, respectively. This can be interpreted as simulating the behavior of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ from the transcript alone: each $a_t$ constructed by Replay equals the one produced by $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ with the same coins $\omega_{\mathcal{M}}$.

The hard cap bounds the number of updates by $K_0 = \big\lfloor 128/(\kappa\epsilon)^2 \big\rfloor + 1$ on every execution. Each entry consists of a round, one of $\mathcal{O}(1/\epsilon)$ rounded answers, and a sign. Hence $\log|\mathcal{Z}| = \mathcal{O}(K_0 \log(T/\epsilon))$. For fixed random coins, each replayed query is independent of $S$, so Hoeffding's inequality and a union bound over $T|\mathcal{Z}|$ queries give the required uniform concentration with

$$M = \mathcal{O}\left(\frac{\log|\mathcal{Z}| + \log(T/\delta)}{\epsilon^2}\right) = \mathcal{O}\big(\rho^2\epsilon^{-4}\log(T/\epsilon) + \epsilon^{-2}\log(1/\delta)\big)\,.$$

The actual query sequence is one of these replays. Lemma B.1 gives the full argument; it uses the hard cap without assuming that the execution is accurate.

**Step 2. $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ outputs $a_t \approx q_t(S)$.** Assume that all empirical means are $\epsilon/8$-close to their population values and all inner-product estimates are $\kappa\epsilon/8$-accurate. On every actual update round, the signed query gives a subgradient of $f_t(x) = \left|\left\langle \tilde{q}_t/\sqrt{N}, x\right\rangle - \kappa q_t(\mathcal{D})\right|$, and $f_t(x_t) > \kappa\epsilon/4$. The surrogate in (2.1) gives a comparator $x_r^\star = \kappa w(g(r))/\sqrt{N}$ with $\|x^\star\| \le 1$ and $f_t(x^\star) \le \kappa\epsilon/16$. By the regret bound for unconstrained OGD, there are at most $128/(\kappa\epsilon)^2$ actual update rounds, so the budget cannot be exhausted. Each answer is therefore either $q_t(S)$ rounded to the nearest multiple of $\epsilon/2$ or $\hat{u}_t/\kappa$ with $|\hat{u}_t/\kappa - q_t(S)| \le \epsilon/2$. Clipping preserves this error bound, and sample concentration then gives population accuracy (Appendix B.1).

## 5 Privacy Guarantee

In this section, we establish the privacy-preserving properties of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$. Although $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ does not satisfy stronger guarantees such as differential privacy [DR14], we show that it enjoys a form of meaningful privacy protection known as Predicate Singling Out (PSO) security [CN20], under the parameter conditions discussed after Theorem 5.2. This notion was introduced as a formal counterpart to the concept of "singling out" that appears in the European Union's General Data Protection Regulation (GDPR).

**Predicate Singling Out (PSO).** The main privacy leakage mode we consider is Predicate Singling Out (PSO). Specifically, given a dataset $S \in \mathcal{I}^M$, a PSO adversary attempts to produce a predicate $p : \mathcal{I} \to \{0,1\}$ that *isolates* a single entry in $S$. Formally, this is captured by the *row isolation* event:

**Definition 5.1** (Row isolation)**.** Let $S = (I_1,\dots,I_M) \in \mathcal{I}^M$ and let $p : \mathcal{I} \to \{0,1\}$. We say that $p$ *isolates* a row in $S$ if there is a unique index $j \in [M]$ such that $p(I_j) = 1$, or equivalently, if $p(S) = 1/M$. We denote this event by $\mathsf{iso}(p,S)$.

We now define PSO security in the context of the adaptive data analysis. Let $\mathcal{P} \subseteq \{0,1\}^{\mathcal{I}}$ be a class of admissible predicates. PSO security is formalized via a game between the ADA mechanism $\mathcal{M}$ and an adversary (or analyst) $\mathcal{A}$, as depicted in Game 2. As in the adaptive query (AQ) game (Game 1), we allow $\mathcal{A}$ to have full power: not only does $\mathcal{A}$ know the data distribution $\mathcal{D}$, but may also choose $\mathcal{D}$ arbitrarily. As before, any guarantee established against such a powerful $\mathcal{A}$ automatically applies to more realistic settings where $\mathcal{D}$ is *unknown* to both $\mathcal{M}$ and $\mathcal{A}$. In fact, the original definition and analysis of PSO security by [CN20] allow for adversaries with this level of power.

**Game 2:** $\mathsf{PSO}_{M,T,\mathcal{P}}[\mathcal{A} \rightleftharpoons \mathcal{M}]$: Predicate Singling Out Game

A multiset $S$ of $M$ i.i.d. samples from $\mathcal{D}$ is given to $\mathcal{M}$ but not to $\mathcal{A}$.
**for** $t \in [T]$ **do**
  $\mathcal{A}$ chooses a query $q_t \in \{-1,1\}^{\mathcal{I}}$ given $(q_1, a_1, \ldots, q_{t-1}, a_{t-1})$.
  $\mathcal{M}$ outputs an answer $a_t \in [-1,1]$ given $(q_1, a_1, \ldots, q_{t-1}, a_{t-1}, q_t)$.
**end**
$\mathcal{A}$ outputs a predicate $p : \mathcal{I} \to \{0,1\}$.

We define the adversary's success probability in the PSO game as:

$$\mathsf{Succ}^{\mathcal{A},\mathcal{M}}_{\mathcal{P},\mathcal{D}}(M,T) := \Pr_{\mathsf{PSO}_{M,T,\mathcal{P}}[\mathcal{A} \rightleftharpoons \mathcal{M}]} \left[ \mathsf{iso}(p,S) \text{ and } p \in \mathcal{P} \right].$$

Following [CN20], fix threshold functions $0 \le w_{\mathsf{low}}(M) \le 1/M \le w_{\mathsf{high}}(M) \le 1$, and define $\mathcal{P}_{\mathsf{low}} = \{p : \mathcal{I} \to \{0,1\} : p(\mathcal{D}) \le w_{\mathsf{low}}(M)\}$ and $\mathcal{P}_{\mathsf{high}} = \{p : \mathcal{I} \to \{0,1\} : p(\mathcal{D}) \ge w_{\mathsf{high}}(M)\}$. We consider $w_{\mathsf{low}} \in \mathsf{negl}(M)$ and $w_{\mathsf{high}} \in \omega(\log M/M)$, where $\mathsf{negl}(M) = \bigcap_{c \in \mathbb{N}} o(M^{-c})$.

For a fixed predicate of mass $\mu = p(\mathcal{D})$ independent of $S$, the isolation probability is $M\mu(1-\mu)^{M-1}$. At $\mu = 1/M$, this approaches $1/e$, so even without any output, a predicate of weight $1/M$ isolates a row with constant probability. For $\mathcal{P}_{\mathsf{low}}$, the baseline is at most $Mw_{\mathsf{low}}(M)$; for $\mathcal{P}_{\mathsf{high}}$, it is at most $Me^{-(M-1)w_{\mathsf{high}}(M)}$. Both bounds are negligible. The low-weight regime concerns exceptionally rare properties, whereas in the high-weight regime exactly one occurrence is unlikely because many records are expected to satisfy the predicate.

We now state the PSO security guarantee of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$, with the full proof deferred to Appendix C.

**Theorem 5.2. [PSO security guarantee]** *Fix $w_{\mathsf{low}} \in \mathsf{negl}(M)$ and $w_{\mathsf{high}} \in \omega(\log M/M)$ as above. For each $\mathcal{P} \in \{\mathcal{P}_{\mathsf{low}}, \mathcal{P}_{\mathsf{high}}\}$, there is a function $\delta_{\mathcal{P}}(M) \in \mathsf{negl}(M)$, depending only on the corresponding threshold, such that for every $\mathcal{D}$ and every adversary $\mathcal{A}$,*

$$\mathsf{Succ}^{\mathcal{A},\mathcal{M}_{\mathsf{AQA\text{-}OGD}}}_{\mathcal{P},\mathcal{D}}(M,T) \le (T/\epsilon)^{\mathcal{O}(\rho^2\epsilon^{-2})}\delta_{\mathcal{P}}(M).$$

*This bound holds for any input norm bound $\rho \ge 1$, without a density assumption on $\mathcal{D}$.*

If the number of possible transcripts is polynomial in $M$, the bound is negligible for both predicate classes. This is the case, for example, when $\rho$ and $\epsilon$ are constant and $T = \mathsf{poly}(M)$. For dense distributions, with $\rho = \lambda^{-1/2}$, the exponent is $\mathcal{O}(\lambda^{-1}\epsilon^{-2})$. The accuracy theorem alone, which permits $M = \mathcal{O}(\log T)$, does not ensure that the number of transcripts is polynomial in $M$. Thus the bound implies negligible PSO success under this additional condition, or when the baseline decays fast enough to offset the number of transcripts. The bound is also quantitatively conservative, because the constant in the exponent is large ($K_0 \ge 513$ when $\epsilon \le 1/2$). Finally, the guarantee excludes all predicate weights strictly between $w_{\mathsf{low}}(M)$ and $w_{\mathsf{high}}(M)$, including $1/M$; for such weights the baseline itself can be non-negligible.

## 6 Discussion and Limitations

Our results show that, with a shared efficient prior sampler, a computationally efficient ADA mechanism attains sample complexity logarithmic in $T$ whenever the target, or a surrogate matching the analyst's queries, has a likelihood ratio with bounded $L_2(\mathcal{D}_g)$ norm. Since the number of lazy OGD updates is capped, the update transcript is always succinct and yields adaptive concentration, the distribution-specific SQ implementation, and the PSO bound. The norm bound ensures that the cap is never reached when the empirical means and inner-product estimates are accurate.

The shared efficient sampler and the surrogate assumption are substantive requirements. A public sample alone does not ensure density outside its support, and pseudodensity depends on the analyst's full query family. It remains open to characterize the weakest condition relative to a prior under which efficient mechanisms retain logarithmic dependence on $T$. There is also a gap between the $\epsilon^{-4}$ upper bound and the $\epsilon^{-2}$ lower bound. Computational efficiency here means polynomial runtime, whose dependence on the accuracy and number of queries may still be large. The PSO result covers specified ranges of predicate weight and yields negligible success under additional parameter conditions. It does not provide differential privacy or unrestricted protection against identification.

## Acknowledgments and Disclosure of Funding

This research received no funding support. The author declares no competing interests.

## References

[AHK12] Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: A meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

[BL13] Maria-Florina Balcan and Phil Long. Active and passive learning of linear separators under log-concave distributions. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pages 288–316. PMLR, 2013.

[Bla25] Guy Blanc. Subsampling suffices for adaptive data analysis. *Journal of the ACM (JACM)*, 72(1):1–45, 2025.

[BLR13] Avrim Blum, Katrina Ligett, and Aaron Roth. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):1–25, 2013.

[BLST25] Guy Blanc, Jane Lange, Carmen Strassle, and Li-Yang Tan. A distributional-lifting theorem for PAC learning. In *Proceedings of the 38th Annual Conference on Learning Theory (COLT)*, pages 375–379. PMLR, 2025.

[BNS+16] Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 1046–1059, 2016.

[BSU17] Mark Bun, Thomas Steinke, and Jonathan Ullman. Make up your mind: The price of online queries in differential privacy. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1306–1325. SIAM, 2017.

[Bub15] Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[CLN+16] Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *Proceedings of the 29th Annual Conference on Learning Theory (COLT)*, pages 772–814. PMLR, 2016.

[CN20] Aloni Cohen and Kobbi Nissim. Towards formalizing the GDPR's notion of singling out. *Proceedings of the National Academy of Sciences*, 117(15):8344–8352, 2020.

[DFH+15a] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *Advances in Neural Information Processing Systems (NeurIPS)*, 28, 2015.

[DFH+15b] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

[DFH+15c] 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 47th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 117–126, 2015.

[DK22] Yuval Dagan and Gil Kur. A bounded-noise mechanism for differential privacy. In *Proceedings of the 35th Annual Conference on Learning Theory (COLT)*, pages 625–661. PMLR, 2022.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[DSU19] Lorenzo De Stefani and Eli Upfal. A Rademacher complexity based method for controlling power and confidence level in adaptive statistical analysis. In *Proceedings of the 6th IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 71–80. IEEE, 2019.

[DSWZ23] Itai Dinur, Uri Stemmer, David P Woodruff, and Samson Zhou. On differential privacy and adaptive data analysis with bounded space. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques (Eurocrypt)*, pages 35–65. Springer, 2023.

[FGV21] Vitaly Feldman, Cristóbal Guzmán, and Santosh Vempala. Statistical Query Algorithms for Mean Vector Estimation and Stochastic Convex Optimization. *Mathematics of Operations Research*, 46(3):912–945, 2021.

[FRR20] Benjamin Fish, Lev Reyzin, and Benjamin IP Rubinstein. Sampling without compromising accuracy in adaptive data analysis. In *Proceedings of 31st International Conference on Algorithmic Learning Theory (ALT)*, pages 297–318. PMLR, 2020.

[FS99] Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

[FS17] Vitaly Feldman and Thomas Steinke. Generalization for adaptively-chosen estimators via stable median. In *Proceedings of the 30th Annual Conference on Learning Theory (COLT)*, pages 728–757. PMLR, 2017.

[FS18] Vitaly Feldman and Thomas Steinke. Calibrating noise to variance in adaptive data analysis. In *Proceedings of the 31st Annual Conference on Learning Theory (COLT)*, pages 535–544. PMLR, 2018.

[GL14] Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102(6):460–465, 2014.

[Haz16] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

[HR10] Moritz Hardt and Guy N Rothblum. A multiplicative weights mechanism for privacy-preserving data analysis. In *Proceedings of IEEE 51st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 61–70. IEEE, 2010.

[HU14] Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *Proceedings of IEEE 55th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 454–463. IEEE, 2014.

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

[JLN+20] Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. In *Proceedings of 11th Innovations in Theoretical Computer Science Conference (ITCS)*, pages 31:1–31:17. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020.

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

[Kha93] Michael Kharitonov. Cryptographic hardness of distribution-specific learning. In *Proceedings of the twenty-fifth annual ACM symposium on Theory of computing*, pages 372–381, 1993.

[KSS22] Aryeh Kontorovich, Menachem Sadigurschi, and Uri Stemmer. Adaptive data analysis with correlated observations. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 11483–11498. PMLR, 2022.

[McM17] H. Brendan McMahan. A Survey of Algorithms and Analysis for Adaptive Online Learning. *Journal of Machine Learning Research*, 18(90):1–50, 2017.

[MR25] Neil G Marchant and Benjamin IP Rubinstein. Adaptive data analysis for growing data. *Advances in Neural Information Processing Systems (NeurIPS)*, 38, 2025.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 94–103. IEEE, 2007.

[NST23] Kobbi Nissim, Uri Stemmer, and Eliad Tsfadia. Adaptive data analysis in a balanced adversarial model. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:25748–25760, 2023.

[RRS+20] Ryan Rogers, Aaron Roth, Adam Smith, Nathan Srebro, Om Thakkar, and Blake Woodworth. Guaranteed validity for empirical approaches to adaptive data analysis. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 2830–2840. PMLR, 2020.

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

[RZ16] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1232–1240. PMLR, 2016.

[Sha18] Ohad Shamir. Distribution-specific hardness of learning neural networks. *Journal of Machine Learning Research*, 19(32):1–29, 2018.

[SL19] Moshe Shenfeld and Katrina Ligett. A necessary and sufficient stability notion for adaptive generalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019.

[SL23] Moshe Shenfeld and Katrina Ligett. Generalization in the face of adaptivity: A Bayesian perspective. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:52153–52165, 2023.

[Smi17] Adam Smith. Information, privacy and stability in adaptive data analysis. *arXiv preprint arXiv:1706.00820*, 2017.

[SU15] 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. PMLR, 2015.

[SU16] Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *Journal of Privacy and Confidentiality*, 7(2):3–22, 2016.

[Ull15] Jonathan Ullman. Private multiplicative weights beyond linear queries. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*, pages 303–312, 2015.

[USN+18] Jonathan Ullman, Adam Smith, Kobbi Nissim, Uri Stemmer, and Thomas Steinke. The limits of post-selection generalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 31:6402–6411, 2018.

[ZH19] Tijana Zrnic and Moritz Hardt. Natural analysts in adaptive data analysis. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 7703–7711. PMLR, 2019.

[Zin03] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 928–936, 2003.

# A Proof of Theorem 3.5

In this appendix, we provide a proof of Theorem 3.5, which states that $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ yields a sample-efficient implementation of a distribution-specific statistical query oracle. For convenience, we restate the theorem below.

**Theorem 3.5. [Sample-efficient DSQ Oracle from $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$]** *Fix a feature distribution $\mathcal{D}_h$ over a finite domain $\mathcal{X}$ generated by an efficient sampler $h$, and an unknown labeling function $f : \mathcal{X} \to \{-1, 1\}$. Let $\mathcal{D}_{h,f}$ denote the induced distribution over $(X, Y)$ given by $X \sim \mathcal{D}_h$ and $Y = f(X)$. Then there exists an implementation of the distribution-specific SQ oracle $\mathsf{DSQ}_h(f, \epsilon)$ with the following guarantee: for any analyst issuing a sequence of $T$ adaptively chosen queries $\phi_1, \ldots, \phi_T$ with $\phi_t : \mathcal{X} \times \{-1, 1\} \to [-1, 1]$, the implementation answers all queries simultaneously to within $\epsilon$ accuracy with probability at least $1 - \delta$, using*

$$M \in \mathcal{O}\big(\epsilon^{-4} \log(BT/\epsilon) + \epsilon^{-2} \log(1/\delta)\big), \quad B = \Theta(\epsilon^{-2} \log(T/\delta))$$

*i.i.d. samples from $\mathcal{D}_{h,f}$, reused for all calls. If $h$ and each $\phi_t$ can be evaluated in $\mathsf{poly}(\log N)$ time, the runtime is $\mathsf{poly}(\log N, \epsilon^{-1}, T, \log(1/\delta))$.*

*Proof of Theorem 3.5.* As discussed above, $\mathcal{D}_{h,f}$ is $(1/2)$-dense with respect to the prior $\mathcal{D}_g$ that samples $X \sim \mathcal{D}_h$ together with an independent uniform label in $\{-1, 1\}$. To handle $[-1, 1]$-valued queries, for each $\phi_t$ draw $B = \lceil 8\epsilon^{-2} \log(4T/\delta) \rceil$ independent thresholds $U_{t,j} \sim \mathrm{Unif}[-1, 1]$ and define

$$q_{t,j}(x, y) = \mathsf{sign}\big(\phi_t(x, y) - U_{t,j}\big).$$

Each threshold is drawn once and held fixed whenever its query is evaluated. Conditional on $\phi_t$ and the history preceding it, the values $q_{t,j}(\mathcal{D}_{h,f})$ are independent, lie in $[-1, 1]$, and have mean $\phi_t(\mathcal{D}_{h,f})$. Hoeffding's inequality bounds the conditional failure probability by $2e^{-B\epsilon^2/8} \le \delta/(2T)$. Averaging over $\phi_t$ and the history and taking a union bound over $t \in [T]$, with probability at least $1 - \delta/2$ we have

$$\left| \frac{1}{B} \sum_{j \in [B]} q_{t,j}(\mathcal{D}_{h,f}) - \phi_t(\mathcal{D}_{h,f}) \right| \le \epsilon/2 \qquad \text{for all } t \in [T]. \tag{A.1}$$

We then run $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ once, using the same fixed labeled sample $S \sim \mathcal{D}_{h,f}^M$, on the resulting sequence of $T' = BT$ Boolean queries with accuracy parameter $\epsilon/2$ and failure probability $\delta/2$, which requires $M \in \mathcal{O}(\epsilon^{-4} \log(BT/\epsilon) + \epsilon^{-2} \log(1/\delta)) = \mathcal{O}(\epsilon^{-4} \log(T/(\epsilon\delta)))$ samples, as $B = \Theta(\epsilon^{-2} \log(T/\delta))$. Drawing the thresholds requires no additional labeled examples. By Corollary 3.3 with $\lambda = 1/2$ (so $\rho = \sqrt{2}$ and $\kappa = 1/\sqrt{2}$), with probability at least $1 - \delta/2$, $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ answers every $q_{t,j}$ to within $\epsilon/2$.

Let $\widehat{\phi}_t$ denote the average of the $B$ answers corresponding to $(q_{t,j})_{j \in [B]}$. This averaging introduces no additional error and approximates $\frac{1}{B} \sum_{j \in [B]} q_{t,j}(\mathcal{D}_{h,f})$ within $\epsilon/2$ error. Combining this with (A.1), we conclude that $|\widehat{\phi}_t - \phi_t(\mathcal{D}_{h,f})| \le \epsilon$ for all $t \in [T]$ with probability at least $1 - \delta$. □

# B Proof of Sample Complexity of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$

In this appendix, we provide a complete proof of our main result, Theorem 3.2, which establishes the sample complexity of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$.

## B.1 Proof of Theorem 3.2

**Step 1. Proving $|q_t(S) - q_t(\mathcal{D})| \in \mathcal{O}(\epsilon)$.** In this step, we prove the following lemma, which establishes the sample complexity required to guarantee uniform concentration between $q_t(S)$ and $q_t(\mathcal{D})$ over all $t \in [T]$.

**Lemma B.1.** [Uniform Concentration] *Let $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ be the ADA mechanism given in Alg. 1. Then, for any adversary $\mathcal{A}$ and $\delta > 0$,*

$$\Pr_{\mathsf{AQ}_{N,M,T}[\mathcal{A} \rightleftharpoons \mathcal{M}_{\mathsf{AQA\text{-}OGD}}]} \Big[ \forall t \in [T],\ |q_t(S) - q_t(\mathcal{D})| \le \frac{\epsilon}{8} \Big] \ge 1 - \frac{\delta}{2},$$

**Algorithm 3:** $\mathsf{Replay}(Z; \mathcal{A}, \omega)$: Reconstruct queries and answers $(q_1, a_1, \ldots, q_T, a_T)$ without samples $S$

**Input:**

- Update transcript $Z \coloneqq \{(t_1, (a_{t_1}, s_{t_1})), \ldots\}$, represented as key-value pairs, where each key $t_i$ denotes a round in which $\mathcal{G}$ was updated, and the corresponding value $(a_{t_i}, s_{t_i})$ records the answer given to $\mathcal{A}$ and the sign of the update at round $t_i$.
- Adversary's algorithm $\mathcal{A}$ and random bits $\omega = (\omega_{\mathcal{A}}, \omega_{\mathcal{M}})$, where $\omega_{\mathcal{A}} \in \{0,1\}^\star$ are used by $\mathcal{A}$, and $\omega_{\mathcal{M}} \in \{0,1\}^\star$ are used by $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ during calls to $\mathsf{IP}_{\mathsf{est}}$ (Alg. 2).

The parameters $g, \rho, \epsilon, \delta, T$ are the same as in Alg. 1.
**Output:** A tuple of reconstructed queries and answers $(q_1, a_1, \ldots, q_T, a_T)$.

$\mathcal{G} \leftarrow \emptyset$
**for** $t \in [T]$ **do**
  $q_t \leftarrow \mathcal{A}^{\omega_{\mathcal{A}}}(q_1, a_1, \ldots, q_{t-1}, a_{t-1})$; set $\tilde{q}_t = q_t \circ g$
    ▷ Run $\mathcal{A}$ with randomness $\omega_{\mathcal{A}}$ and previous query-answer history.
  $\hat{u}_t \leftarrow \mathsf{IP}_{\mathsf{est}}^{\omega_{\mathcal{M}}}(\tilde{q}_t, \mathcal{G}; \kappa\epsilon/8, \delta/2T)$
    ▷ Run $\mathsf{IP}_{\mathsf{est}}$ using $\omega_{\mathcal{M}}$, reading from where the last $\mathsf{IP}_{\mathsf{est}}$ call left off.
  **if** $t \in Z.\mathsf{Keys}$ **then**
    $(a_t, s_t) \leftarrow Z.\mathsf{Get}(t)$ ▷ Get the entry with key $t$.
    $\mathcal{G} \leftarrow \mathcal{G} \cup \{-\kappa\epsilon s_t \tilde{q}_t/16\}$ ▷ Update $\mathcal{G}$.
    $Z.\mathsf{Remove}(t)$ ▷ Remove the entry with key $t$.
  **end**
  **else**
    $a_t \leftarrow \min\{1, \max\{-1, \hat{u}_t/\kappa\}\}$
  **end**
**end**
**return** $(q_1, a_1, \ldots, q_T, a_T)$

*whenever* $M \geq C\big(\rho^2\epsilon^{-4}\log(T/\epsilon) + \epsilon^{-2}\log(1/\delta)\big)$ *for a universal constant* $C$.

*Proof of Lemma B.1.* We prove the lemma by explicitly constructing an algorithm, Replay given in Alg. 3, that reconstructs queries $(q_1, \ldots, q_T)$ asked by $\mathcal{A}$ during game $\mathsf{AQ}_{N,M,T}[\mathcal{A} \rightleftharpoons \mathcal{M}_{\mathsf{AQA\text{-}OGD}}]$, from a succinct transcript.

A main input to Replay is the update transcript, $Z \coloneqq \{(t_1, (a_{t_1}, s_{t_1})), \ldots\}$ represented by a set of key-value pairs $(t_i, (a_{t_i}, s_{t_i}))$ where the key $t_i \in [T]$ is a round in which $\mathcal{G}$ was updated and $(a_{t_i}, s_{t_i})$ is the pair of the answer provided to $\mathcal{A}$ and the sign of the update, $\mathsf{sign}(\hat{u}_{t_i} - \kappa q_{t_i}(S))$, at round $t_i$.

Since the if-statement of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ allows an update only while $|\mathcal{G}| \leq 128(\kappa\epsilon)^{-2}$, $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ updates $\mathcal{G}$ at most $K_0 \coloneqq \lfloor 128(\kappa\epsilon)^{-2} \rfloor + 1$ times, and the number of entries in $Z$ is at most $K_0$. Each entry consists of a round in $[T]$, an answer $a_t$, which is $q_t(S) \in [-1,1]$ rounded to a multiple of $\epsilon/2$ and clipped to $[-1,1]$ and hence takes at most $4/\epsilon + 2$ values, and a sign $s_t \in \{\pm 1\}$. Padding $Z$ with empty entries to $K_0$ entries, there are at most $\big((T+1)(8/\epsilon+4)\big)^{K_0} \leq (24T/\epsilon)^{K_0}$ possible choices for $Z$, using $\epsilon \leq 1$. Let the set of all such logs be denoted by $\mathcal{Z}$. The hard cap bounds this set on every execution, including those on which accuracy fails.

Then, for any fixed random bits $\omega$ used by $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ and $\mathcal{A}$ (including those used by $\mathcal{A}$ to select the distribution $\mathcal{D}$), we have

$$\Pr_{S\sim\mathcal{D}^M}\left[\forall(t,Z) \in [T]\times\mathcal{Z},\ |q_t(S) - q_t(\mathcal{D})| \leq \sqrt{\frac{2\big(\log|\mathcal{Z}| + \log(4T/\delta)\big)}{M}},\right.$$
$$\left.\text{where } (q_t, a_t)_{t\in[T]} \leftarrow \mathsf{Replay}(Z; \mathcal{A}, \omega)\right] \geq 1 - \frac{\delta}{2} \quad \text{(B.1)}$$

by applying the union bound over $[T] \times \mathcal{Z}$ to the Hoeffding bound $\Pr_{S\sim\mathcal{D}^M}[|q(S) - q(\mathcal{D})| > \tau] \leq 2e^{-M\tau^2/2}$, which holds for each fixed query $q \in \{-1, 1\}^{\mathcal{I}}$. The right-hand side of the inequality in (B.1) is at most $\epsilon/8$ whenever $M \geq 128\epsilon^{-2}\big(\log|\mathcal{Z}| + \log(4T/\delta)\big)$, which holds under the condition of the lemma, as $\log|\mathcal{Z}| \leq K_0 \log(24T/\epsilon)$.

For each randomness $\omega$ and sample $S$, there is $Z(\omega, S) \in \mathcal{Z}$ such that $\mathsf{Replay}(Z(\omega, S); \mathcal{A}, \omega)$ reconstructs the queries $(q_t)_{t\in[T]}$ asked in game $\mathsf{AQ}_{N,M,T}[\mathcal{A} \rightleftharpoons \mathcal{M}_{\mathsf{AQA\text{-}OGD}}]$ when run with fixed $\omega$ and $S$. Hence, taking the expectation over $\omega$ in (B.1), this fact yields the bound stated in the lemma. □

**Step 2. Finalizing the proof by showing $|a_t - q_t(S)| \in \mathcal{O}(\epsilon)$.** With the high-probability concentration bound $|q_t(S) - q_t(\mathcal{D})| \in \mathcal{O}(\epsilon)$ established in Step 1, we now complete the proof of Theorem 3.2 by showing that $|a_t - q_t(S)| \in \mathcal{O}(\epsilon)$ for all $t \in [T]$.

This follows from the fact that the number of updates made by Alg. 1 remains small, i.e., $|\mathcal{G}| \leq 128/(\kappa\epsilon)^2$, as ensured by a regret argument. In particular, we obtain a contradiction by showing that exceeding this number of updates would violate the regret bound for OGD. Consequently, the condition $|\mathcal{G}| \leq 128/(\kappa\epsilon)^2$ in the if-statement of Alg. 1 is redundant, and $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ therefore outputs approximately correct answers.

Before proceeding, we formally state the regret bound for unconstrained OGD. While this result is standard (see, e.g., [Bub15, Haz16, Zin03]), we include a self-contained proof in Appendix B.2 for completeness.

**Lemma B.2.** [Regret Bound for Unconstrained OGD] *For any finite ordered set $\mathcal{T} := \{t_1, \ldots, t_{|\mathcal{T}|}\} \subset \mathbb{N}$ such that $t_1 < \cdots < t_{|\mathcal{T}|}$, and for each $t \in \mathcal{T}$, let $f_t : \mathbb{R}^N \to \mathbb{R}$ be any convex loss function with unit-bounded subgradients, i.e., $\|\nabla f_t\| \leq 1$. Then, for any $\alpha > 0$ and $x^\star \in \mathbb{B}^N$,*

$$\sum_{t\in\mathcal{T}} \big[f_t(x_t) - f_t(x^\star)\big] \leq \frac{1}{2\alpha} + \frac{\alpha|\mathcal{T}|}{2}, \tag{B.2}$$

*where $x_t$ is defined recursively as*

$$x_{t_{i+1}} \coloneqq x_{t_i} - \alpha\nabla f_{t_i}(x_{t_i}) \quad \text{and} \quad x_{t_1} \coloneqq 0. \tag{B.3}$$

**FTRL interpretation.** Let $\mathcal{T}_{\text{mis}} \subseteq [T]$ denote the rounds in which $\mathcal{G}$ is actually updated, and let $\gamma_u = s_u\tilde{q}_u/\sqrt{N}$ for $u \in \mathcal{T}_{\text{mis}}$. With $\alpha = \kappa\epsilon/16$, the normalized sketch $x_t$ at round $t$ satisfies the identity

$$x_t = -\alpha \sum_{u\in\mathcal{T}_{\text{mis}},\, u<t} \gamma_u = \operatorname*{arg\,min}_{x\in\mathbb{R}^N} \left\{ \sum_{u\in\mathcal{T}_{\text{mis}},\, u<t} \langle\gamma_u, x\rangle + \frac{\|x\|^2}{2\alpha} \right\}. \tag{B.4}$$

This follows by differentiating the quadratic objective and is the standard constant-step OGD/FTRL equivalence [McM17, Section 3.1]. The identity holds for the recorded signs on every execution. In the proof below, accurate inner-product estimates and sample concentration ensure that these signs give subgradients of the population absolute losses on update rounds. Thus the FTRL objective accumulates their linearizations; no constraint or projection onto $\mathbb{B}^N$ is imposed on the iterates.

With all the necessary ingredients established so far, we now present the proof of Theorem 3.2, which we restate below for convenience.

**Theorem 3.2. [Sample Efficiency]** *If $\mathcal{A}$ almost surely chooses a distribution $\mathcal{D}$ for which some surrogate $\mathcal{D}'$ satisfies* (2.1) *with respect to its query family $\mathcal{Q}^{\mathcal{A}}$, then $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$, run with input $\rho$, is $(\epsilon, \delta)$-accurate with respect to $\mathcal{A}$, with any sample size $M \geq C\big(\rho^2\epsilon^{-4}\log(T/\epsilon) + \epsilon^{-2}\log(1/\delta)\big)$, where $C$ is a universal constant.*

*Proof of Theorem 3.2.* Recall that $\mathcal{T}_{\text{mis}} \subseteq [T]$ denotes the rounds in which $\mathcal{G}$ is actually updated. We first bound the size of this set using OGD on these rounds only. If $|\mathcal{T}_{\text{mis}}| \leq 128/(\kappa\epsilon)^2$, the update budget cannot be exhausted: reaching a state above the cap would require $\lfloor 128/(\kappa\epsilon)^2 \rfloor + 1$ actual updates. Thus every threshold violation causes an update, and every other round has $|\hat{u}_t/\kappa -$

$q_t(S)| \le \epsilon/2$. Rounding introduces at most $\epsilon/4$ error and clipping cannot increase error relative to $q_t(S) \in [-1, 1]$.

Let $\mathsf{E}$ be the event from Lemma B.1, that is $\forall t \in [T],\ |q_t(S) - q_t(\mathcal{D})| \le \epsilon/8$, and $\mathsf{F}$ be the event that $\mathsf{IP}_{\mathsf{est}}$ produces a $\kappa\epsilon/8$-accurate estimate for all $t \in [T]$. We show that when both events occur (i.e., $\mathsf{E} \cap \mathsf{F}$), we have $|\mathcal{T}_{\text{mis}}| \le \frac{128}{(\kappa\epsilon)^2}$, hence $|a_t - q_t(S)| \le \epsilon/2$ for all $t \in [T]$. Moreover, under $\mathsf{E}$ we have $|q_t(S) - q_t(\mathcal{D})| \le \epsilon/8$ for all $t \in [T]$. Hence, under $\mathsf{E} \cap \mathsf{F}$, we have $|a_t - q_t(\mathcal{D})| \le \epsilon$ for all $t \in [T]$, which is the theorem's claim.

Therefore, it suffices to show the following two claims:

**Claim 1.** $\Pr_{\mathsf{AQ}_{N,M,T}[\mathcal{A} \rightleftharpoons \mathcal{M}_{\mathsf{AQA\text{-}OGD}}]}[\mathsf{E} \cap \mathsf{F}] \ge 1 - \delta$.

**Claim 2.** Given that $\mathsf{E} \cap \mathsf{F}$ occurs and $\mathcal{D}$ admits a surrogate satisfying (2.1), $|\mathcal{T}_{\text{mis}}| \le \frac{128}{(\kappa\epsilon)^2}$.

*Proof of Claim 1.* Lemma B.1 shows that $\Pr_{\mathsf{AQ}_{N,M,T}[\mathcal{A} \rightleftharpoons \mathcal{M}_{\mathsf{AQA\text{-}OGD}}]}[\mathsf{E}] \ge 1 - \delta/2$. Hence, it is enough to show that

$$\mathsf{F} \coloneqq \left\{ \forall t \in [T],\ \left| \hat{u}_t - \frac{1}{N} \langle \tilde{q}_t, y_t \rangle \right| \le \frac{\kappa\epsilon}{8} \right\},$$

occurs with probability at least $1 - \delta/2$ under $\mathsf{AQ}_{N,M,T}[\mathcal{A} \rightleftharpoons \mathcal{M}_{\mathsf{AQA\text{-}OGD}}]$, where $\hat{u}_t$ is the output of $\mathsf{IP}_{\mathsf{est}}(\tilde{q}_t, \mathcal{G}; \kappa\epsilon/8, \delta/2T)$ and $y_t \coloneqq \sum_{v \in \mathcal{G}} v$ for the multiset $\mathcal{G}$ at the start of round $t$. Since the randomness used by $\mathsf{IP}_{\mathsf{est}}$ to estimate $\langle \tilde{q}_t, y_t \rangle / N$ is independent from $q_t$ and $\mathcal{G}$, we can apply the union bound over Lemma B.3, which yields

$$\Pr_{\mathsf{AQ}_{N,M,T}[\mathcal{A} \rightleftharpoons \mathcal{M}_{\mathsf{AQA\text{-}OGD}}]}[\,\mathsf{F}\,] \ge 1 - \frac{\delta}{2}.$$

□

*Proof of Claim 2.* For each $t \in \mathcal{T}_{\text{mis}}$, define $f_t : \mathbb{R}^N \to \mathbb{R}$ as

$$f_t(x) \coloneqq |h_t(x)|\,, \quad h_t(x) \coloneqq \left\langle \frac{\tilde{q}_t}{\sqrt{N}}, x \right\rangle - \kappa q_t(\mathcal{D}).$$

so that $f_t$ is convex, and its sub-gradient satisfies $\|\nabla f_t(x)\| = \|\mathsf{sign}(h_t(x))\tilde{q}_t/\sqrt{N}\| = 1$, since $\tilde{q}_t \in \{\pm 1\}^N$.

For each $t \in \mathcal{T}_{\text{mis}}$, let $x_t \coloneqq y_t/\sqrt{N}$, where $y_t \coloneqq \sum_{v \in \mathcal{G}} v$ is computed using the gradient set $\mathcal{G}$ at round $t$. Notice that, by construction, $\mathcal{G}$ consists of vectors $-\kappa\epsilon s_t \tilde{q}_t/16$, where $s_t \coloneqq \mathsf{sign}(\hat{u}_t - \kappa q_t(S))$ and $|\hat{u}_t - \kappa q_t(S)| > \kappa\epsilon/2$. Now, we list two key implications under $\mathsf{E} \cap \mathsf{F}$:

- Given that $\mathsf{E} \cap \mathsf{F}$ has occurred, $|\hat{u}_t - \kappa q_t(S)| > \kappa\epsilon/2$ implies

  $$\mathsf{sign}(\hat{u}_t - \kappa q_t(S)) = \mathsf{sign}\left( \left\langle \frac{\tilde{q}_t}{\sqrt{N}}, x_t \right\rangle - \kappa q_t(\mathcal{D}) \right) = \mathsf{sign}(h_t(x_t)),$$

  as $|\hat{u}_t - \langle \tilde{q}_t, y_t \rangle / N| = \left| \hat{u}_t - \left\langle \tilde{q}_t/\sqrt{N}, x_t \right\rangle \right| \le \kappa\epsilon/8$ under $\mathsf{F}$ and $|q_t(S) - q_t(\mathcal{D})| \le \epsilon/8$ under $\mathsf{E}$.

  Hence, $\mathcal{G}$ consists of vectors of the form $-\alpha\sqrt{N}\nabla f_t(x_t)$ with $\alpha = \kappa\epsilon/16$, and thus $x_t \coloneqq \sum_{v \in \mathcal{G}} v/\sqrt{N}$ follows the OGD update rule given in (B.3), with learning rate $\kappa\epsilon/16$.

- Similarly, under $\mathsf{E} \cap \mathsf{F}$, $|\hat{u}_t - \kappa q_t(S)| > \kappa\epsilon/2$ implies $f_t(x_t) > \kappa\epsilon/4$.

From the above implications together with the OGD regret bound (B.2), we observe the following: For any $x^\star \in \mathbb{B}^N$,

$$\frac{\kappa\epsilon}{4}|\mathcal{T}_{\text{mis}}| - \sum_{t \in \mathcal{T}_{\text{mis}}} f_t(x^\star) \le \sum_{t \in \mathcal{T}_{\text{mis}}} \left[ f_t(x_t) - f_t(x^\star) \right] \le \frac{8}{\kappa\epsilon} + \frac{\kappa\epsilon}{32}|\mathcal{T}_{\text{mis}}|$$

$$\implies \frac{\kappa\epsilon|\mathcal{T}_{\text{mis}}|}{8} - \sum_{t \in \mathcal{T}_{\text{mis}}} f_t(x^\star) \le \frac{8}{\kappa\epsilon}.$$

Hence, provided that there is some $x^\star \in \mathbb{B}^N$ such that

$$\sum_{t \in \mathcal{T}_{\text{mis}}} f_t(x^\star) \le \frac{\kappa\epsilon|\mathcal{T}_{\text{mis}}|}{16},$$

we obtain the desired upper bound on the mistake set: $|\mathcal{T}_{\text{mis}}| \le 128/(\kappa\epsilon)^2$. Indeed, this follows from (2.1), as we show below:

**Claim 3.** Under (2.1), there is $x^\star \in \mathbb{B}^N$ such that $\sum_{t \in \mathcal{T}_{\text{mis}}} f_t(x^\star) \le \kappa\epsilon|\mathcal{T}_{\text{mis}}|/16$.

*Proof of Claim 3.* Choose the surrogate $\mathcal{D}'$ from (2.1), and let $w = d\mathcal{D}'/d\mathcal{D}_g$. Define the comparator in seed space by

$$x^\star_r = \frac{\kappa w(g(r))}{\sqrt{N}}, \qquad r \in [N].$$

Then

$$\|x^\star\|^2 = \frac{\kappa^2}{N}\sum_{r=1}^{N} w(g(r))^2 = \kappa^2 \mathop{\mathbb{E}}_{\mathcal{D}_g}[w^2] \le 1,$$

$$\left\langle \frac{\tilde{q}_t}{\sqrt{N}}, x^\star \right\rangle = \kappa \mathop{\mathbb{E}}_{I \sim \mathcal{D}_g}[q_t(I)w(I)] = \kappa q_t(\mathcal{D}').$$

Since every issued query belongs to $\mathcal{Q}^{\mathcal{A}}$, we have $f_t(x^\star) = \kappa|q_t(\mathcal{D}') - q_t(\mathcal{D})| \le \kappa\epsilon/16$. Summing over the actual update rounds proves the claim. The comparator is used only in the analysis and need not be computed by the mechanism. □

Therefore, we conclude the proof of Claim 2. □

Claims 1 and 2 complete the proof of Theorem 3.2. □

### B.2 Proofs for Technical Lemmas

In this subsection, we collect technical lemmas and their proofs used in the proof of Theorem 3.2. We first restate, for convenience, the regret bound for unconstrained OGD and provide its proof [Zin03] for completeness.

**Lemma B.2.** [Regret Bound for Unconstrained OGD] *For any finite ordered set $\mathcal{T} \coloneqq \{t_1, \dots, t_{|\mathcal{T}|}\} \subset \mathbb{N}$ such that $t_1 < \cdots < t_{|\mathcal{T}|}$, and for each $t \in \mathcal{T}$, let $f_t : \mathbb{R}^N \to \mathbb{R}$ be any convex loss function with unit-bounded subgradients, i.e., $\|\nabla f_t\| \le 1$. Then, for any $\alpha > 0$ and $x^\star \in \mathbb{B}^N$,*

$$\sum_{t \in \mathcal{T}} \big[f_t(x_t) - f_t(x^\star)\big] \le \frac{1}{2\alpha} + \frac{\alpha|\mathcal{T}|}{2}, \tag{B.2}$$

*where $x_t$ is defined recursively as*

$$x_{t_{i+1}} \coloneqq x_{t_i} - \alpha\nabla f_{t_i}(x_{t_i}) \quad \text{and} \quad x_{t_1} \coloneqq 0. \tag{B.3}$$

*Proof of Lemma B.2.* Fix any $x^\star \in \mathbb{B}^N$. We begin by observing that

$$\begin{aligned}\|x_{t_{i+1}} - x^\star\|^2 &= \|x_{t_i} - \alpha\nabla f_{t_i}(x_{t_i}) - x^\star\|^2 \\ &= \|x_{t_i} - x^\star\|^2 - 2\alpha\langle\nabla f_{t_i}(x_{t_i}), x_{t_i} - x^\star\rangle + \alpha^2\|\nabla f_{t_i}(x_{t_i})\|^2.\end{aligned}$$

As $\|\nabla f_{t_i}(x_{t_i})\| \le 1$, rearranging the above, we have that

$$\langle\nabla f_{t_i}(x_{t_i}), x_{t_i} - x^\star\rangle \le \frac{\|x_{t_i} - x^\star\|^2 - \|x_{t_{i+1}} - x^\star\|^2}{2\alpha} + \frac{\alpha}{2}.$$

Since $f_t$s are convex, $f_{t_i}(x_{t_i}) - f_{t_i}(x^\star) \le \langle\nabla f_{t_i}(x_{t_i}), x_{t_i} - x^\star\rangle$ for all $i$. Thus summing the above over all $i = 1, \dots, |\mathcal{T}|$,

$$\sum_{t \in \mathcal{T}} f_t(x_t) - f_t(x^\star) \le \sum_{i=1}^{|\mathcal{T}|} \langle\nabla f_{t_i}(x_{t_i}), x_{t_i} - x^\star\rangle$$

$$\leq \frac{\|x_{t_1} - x^\star\|^2}{2\alpha} + \frac{\alpha|\mathcal{T}|}{2}$$
$$\leq \frac{1}{2\alpha} + \frac{\alpha|\mathcal{T}|}{2},$$

as $x_{t_1} = 0$. □

Finally, we provide a simple proof of correctness for the subroutine $\mathsf{IP}_{\mathsf{est}}$.

**Lemma B.3.** [Correctness of $\mathsf{IP}_{\mathsf{est}}$] *Suppose $|\mathcal{G}| \leq T$ and every $v \in \mathcal{G}$ has the form $v = \alpha q'$ with $|\alpha| \leq 1$ and $q' \in \{-1, 1\}^N$, as in Alg. 1. The output of $\mathsf{IP}_{\mathsf{est}}(q, \mathcal{G}; \epsilon, \delta)$ (Alg. 2) lies in the interval $\left[\langle q, y\rangle / N - \epsilon, \langle q, y\rangle / N + \epsilon\right]$ where $y \coloneqq \sum_{v\in\mathcal{G}} v$, with probability at least $1 - \delta$.*

*Proof of Lemma B.3.* The estimand is

$$\frac{1}{N}\langle q, v\rangle = \mathop{\mathbb{E}}_{i\sim U_{[N]}}[q_i v_i].$$

Since every vector $v \in \mathcal{G}$ can be written as $v = \alpha q'$ for some $|\alpha| \leq 1$ and $q' \in \{-1, 1\}^N$, and $q \in \{-1, 1\}^N$, we have $|q_i v_i| \leq 1$ for all $i \in [N]$. With the stated sample size, Hoeffding's inequality bounds the estimation error for each $v$ by $\epsilon/T$ except with probability $\delta/T$. A union bound over at most $T$ vectors and summation of these errors gives total error at most $\epsilon$ with probability at least $1 - \delta$. The case $\mathcal{G} = \emptyset$ is exact. □

## C Proof of Theorem 5.2

In this appendix, we provide a proof of Theorem 5.2, which establishes the PSO security guarantee of $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$. For convenience, we restate the theorem below.

**Theorem 5.2. [PSO security guarantee]** *Fix $w_{\mathsf{low}} \in \mathsf{negl}(M)$ and $w_{\mathsf{high}} \in \omega(\log M/M)$ as above. For each $\mathcal{P} \in \{\mathcal{P}_{\mathsf{low}}, \mathcal{P}_{\mathsf{high}}\}$, there is a function $\delta_{\mathcal{P}}(M) \in \mathsf{negl}(M)$, depending only on the corresponding threshold, such that for every $\mathcal{D}$ and every adversary $\mathcal{A}$,*

$$\mathsf{Succ}^{\mathcal{A},\mathcal{M}_{\mathsf{AQA\text{-}OGD}}}_{\mathcal{P},\mathcal{D}}(M, T) \leq (T/\epsilon)^{\mathcal{O}(\rho^2\epsilon^{-2})}\delta_{\mathcal{P}}(M).$$

*This bound holds for any input norm bound $\rho \geq 1$, without a density assumption on $\mathcal{D}$.*

*Proof of Theorem 5.2.* Fix the random strings used by $\mathcal{A}$ and $\mathcal{M}$, denoted by $\omega_{\mathcal{A}}$ and $\omega_{\mathcal{M}}$ respectively, and let $\omega \coloneqq (\omega_{\mathcal{A}}, \omega_{\mathcal{M}})$. Then, we may use $\mathsf{Replay}(Z; \mathcal{A}, \omega)$ from Alg. 3 to simulate the interaction between $\mathcal{A}$ and $\mathcal{M}_{\mathsf{AQA\text{-}OGD}}$ and recover the predicate $p = \mathcal{A}^{\omega_{\mathcal{A}}}(q_1, a_1, \ldots, q_T, a_T)$ that $\mathcal{A}$ outputs. By the same argument used in the proof of Lemma B.1, there exists a transcript $Z(\omega, S) \in \mathcal{Z}$ such that $\mathcal{A}^{\omega_{\mathcal{A}}}\big(\mathsf{Replay}(Z(\omega, S); \mathcal{A}, \omega)\big)$ outputs the correct predicate $p$. Furthermore, the total number of possible transcripts satisfies $|\mathcal{Z}| \leq (24T/\epsilon)^{K_0}$ with $K_0 = \left\lfloor 128(\kappa\epsilon)^{-2}\right\rfloor + 1$, with $\kappa = 1/\rho$ (see the proof of Lemma B.1 in Appendix B.1).

For fixed $\omega$ and $Z$, the predicate $p_Z = \mathcal{A}^{\omega_{\mathcal{A}}}(\mathsf{Replay}(Z; \mathcal{A}, \omega))$ is independent of $S$. Define

$$\mathrm{base}_{\mathcal{P},\mathcal{D}}(M) = \sup_{p\in\mathcal{P}} Mp(\mathcal{D})(1 - p(\mathcal{D}))^{M-1},$$

with value zero when $\mathcal{P}$ is empty. A union bound over all transcripts gives

$$\Pr_{S\sim\mathcal{D}^M}[\mathsf{iso}(p_{Z(\omega,S)}, S),\ p_{Z(\omega,S)} \in \mathcal{P}] \leq \sum_{Z\in\mathcal{Z}} \Pr_{S\sim\mathcal{D}^M}[\mathsf{iso}(p_Z, S),\ p_Z \in \mathcal{P}]$$
$$\leq |\mathcal{Z}|\,\mathrm{base}_{\mathcal{P},\mathcal{D}}(M).$$

Averaging over $\omega$ yields

$$\mathsf{Succ}^{\mathcal{A},\mathcal{M}_{\mathsf{AQA\text{-}OGD}}}_{\mathcal{P},\mathcal{D}}(M, T) \leq \min\{1, |\mathcal{Z}|\,\mathrm{base}_{\mathcal{P},\mathcal{D}}(M)\}.$$

For the fixed threshold families in §5,

$$\mathrm{base}_{\mathcal{P}_{\mathsf{low}},\mathcal{D}}(M) \leq Mw_{\mathsf{low}}(M),$$

$$\mathrm{base}_{\mathcal{P}_{\mathsf{high}},\mathcal{D}}(M) \le M e^{-(M-1)w_{\mathsf{high}}(M)}.$$

Write $\mu = p(\mathcal{D})$. The first bound follows by dropping $(1-\mu)^{M-1} \le 1$. For the second, use $\mu \le 1$, $1-\mu \le e^{-\mu}$, and $\mu \ge w_{\mathsf{high}}(M)$. Both bounds are negligible under the specified threshold assumptions, uniformly over $\mathcal{D}$. Taking these as $\delta_{\mathcal{P}}(M)$ and using $|\mathcal{Z}| \le (24T/\epsilon)^{K_0}$ proves the theorem. Only the hard update cap is needed; neither sample accuracy nor the surrogate assumption is used.

□

# D A Lower Bound for Constant-Density Distributions

**Proposition 3.4. [Lower Bound]** *For $T \ge 2$ and $0 < \epsilon \le 1/8$, there is a family of $2/3$-dense distributions relative to the uniform prior on $\{-1,1\}^T$ such that answering the $T$ fixed coordinate queries simultaneously within $\epsilon$, with probability at least $2/3$, requires $M = \Omega(\epsilon^{-2}\log T)$ samples.*

*Proof.* Let $U$ be uniform on $\{-1,1\}^T$. For each $j \in [T]$ and $s \in \{-1,1\}$, define

$$\mathcal{D}_{j,s}(x) = 2^{-T}(1 + 4\epsilon s x_j).$$

Since $\mathbb{E}_U\, x_j = 0$, this is normalized. For $\epsilon \le 1/8$, its likelihood ratio lies in $[1/2, 3/2]$, so it is nonnegative and $2/3$-dense relative to $U$. For the fixed coordinate queries $q_k(x) = x_k$,

$$q_k(\mathcal{D}_{j,s}) = 4\epsilon s\, \mathbf{1}\{k = j\}.$$

If all answers have error at most $\epsilon$, the unique coordinate whose answer has magnitude greater than $2\epsilon$ identifies $j$, and its sign identifies $s$.

Let $\Theta = (j,s)$ be uniform on the $2T$ choices, and draw $S \sim \mathcal{D}_\Theta^M$. Writing $a = 4\epsilon$ and using natural logarithms,

$$\begin{aligned} D_{\mathrm{KL}}(\mathcal{D}_{j,s}\|U) &= \frac{(1+a)\log(1+a) + (1-a)\log(1-a)}{2} \\ &\le \frac{(1+a)a - (1-a)a}{2} = a^2 = 16\epsilon^2, \end{aligned}$$

where the inequality uses $\log z \le z - 1$. If $\overline{P}$ denotes the mixture of the laws $\mathcal{D}_\Theta^M$, the relative-entropy identity gives

$$I(\Theta; S) = \mathop{\mathbb{E}}_{\Theta} D_{\mathrm{KL}}(\mathcal{D}_\Theta^M \| U^M) - D_{\mathrm{KL}}(\overline{P}\|U^M) \le 16M\epsilon^2.$$

The mechanism's random coins are independent of $(\Theta, S)$. Its answers therefore give an estimator of $\Theta$ with error probability at most $1/3$. Fano's inequality yields

$$16M\epsilon^2 \ge \frac{2}{3}\log(2T) - \log 2 \ge \frac{1}{3}\log T,$$

where the last inequality uses $T \ge 2$. Hence $M \ge (48\epsilon^2)^{-1}\log T$. This construction uses a domain of size $2^T$. Since the upper bound of Corollary 3.3 holds for every domain size, its dependence on $T$ cannot be improved; the gap between the upper bound's $\epsilon^{-4}$ and the lower bound's $\epsilon^{-2}$ dependence remains. □

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's main claims, presented in the abstract and introduction, summarize the paper's contributions and scope. Moreover, we provide an Organization subsection to guide readers to the relevant sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses the shared sampler and surrogate assumptions, the remaining accuracy gap, and the restrictions on the PSO guarantee. The precise assumptions appear in Sections 2 and 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We list our assumptions in the preliminary section, and each theorem and lemma explicitly states the assumptions required for its application. The main accuracy assumption is stated in (2.1), with pointwise density as a corollary. Proofs are in Appendices A–D.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [N/A]

Justification: Because our contribution is theoretical, the paper does not include experimental results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: Our paper does not include experiments requiring code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We reviewed the NeurIPS Code of Ethics and confirmed that this paper conforms to it in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: This theoretical work does not evaluate societal impacts of particular deployments. Sections 5 and 6 explain the scope of the PSO bound and its limitations for privacy protection.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper's contribution is theoretical and there is no such risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.