Interaction is necessary for distributed learning with privacy or communication constraints

Yuval Dagan* MIT

Vitaly Feldman Google Research

Abstract

Local differential privacy (LDP) is a model where users send privatized data to an untrusted central server whose goal it to solve some data analysis task. In the non-interactive version of this model the protocol consists of a single round in which a server sends requests to all users then receives their responses. This version is deployed in industry due to its practical advantages and has attracted significant research interest.

Our main result is an exponential lower bound on the number of samples necessary to solve the standard task of learning a large-margin linear separator in the non-interactive LDP model. Via a standard reduction this lower bound implies an exponential lower bound for stochastic convex optimization and specifically, for learning linear models with a convex, Lipschitz and smooth loss. These results answer the questions posed in [STU17; DF18]. Our lower bound relies on a new technique for constructing pairs of distributions with nearly matching moments but whose supports can be nearly separated by a large margin hyperplane. These lower bounds also hold in the model where communication from each user is limited and follow from a lower bound on learning using non-adaptive *statistical queries*.

1 Introduction

The primary model we study is distributed learning with the constraint of local differential privacy (LDP) [War65; EGS03; Kas+11]. In this model each client (or user) holds an individual data point and a server can communicate with the clients. The goal of the server is to solve some statistical analysis on the data stored at the clients. In addition, the server is not trusted and the communication should not reveal significant private information about the users' data. Specifically, the entire protocol needs to satisfy differential privacy [Dwo+06]. In the general version of the model, the executed protocol can involve an arbitrary number of rounds of interaction between the server and the clients. In practice, however, network latencies significantly limit the number of rounds of interaction that can be executed. Indeed, currently deployed systems that use local differential privacy are non-interactive [EPK14; App17; DKY17]. Namely, the server sends each client a request; based on the request each client runs some differentially private algorithm on its data and sends a response back to the server. The server then analyzes the data it received (without further communication with the clients). See Section 2.1 for a formal definition of the model.

This motivates the question: which problems can be solved by non-interactive LDP protocols? This question was first formally addressed by Kasiviswanathan, Lee, Nissim, Raskhodnikova, and Smith [Kas+11] who also established an equivalence, up to polynomial factors, between algorithms in the statistical query (SQ) framework of Kearns [Kea98] and LDP protocols[1]. In this equivalence, non-interactive protocols correspond to non-adaptive SQ algorithms. Unfortunately, most SQ learning algorithms are adaptive and thus, for most problems, this equivalence only gives interactive LDP protocols. Using this equivalence, Kasiviswanathan

*Part of the work was done while the author was at Google Research.

[1]More formally, the equivalence is for a more restricted way to measure privacy based on composition of the privacy parameters of each message sent by a user.

et al. [Kas+11] also constructed an artificial learning problem which requires an exponentially larger number of samples to solve by any non-interactive LDP protocol than it does when interaction is allowed.

Motivated by the industrial applications of the LDP model, Smith, Thakurta, and Upadhyay [STU17] studied the complexity of solving stochastic convex loss minimization problems by non-interactive LDP algorithms. In these problems we are given a family of loss functions $\{\ell(w;z)\}_{z\in Z}$ convex in w and a convex body $\mathcal{K}\subseteq\mathbb{R}^d$. For a distribution P over Z the goal is to find an approximate minimizer of

$$\ell(w;P) := \mathop{\mathbb{E}}_{z\sim P}\ell(w;z).$$

over $w\in\mathcal{K}$. They gave a non-interactive LDP algorithm that uses an exponential in d number of samples. Additionally, they showed that such dependence is unavoidable for the commonly used optimization algorithms whose queries rely solely on the information in the neighborhood of the query point w (such as gradients or Hessians). Their bounds have been strengthened and generalized in a number of subsequent works [DRY18; Woo+18; BS18; DG18; WGX18] but the question of whether a non-interactive LDP protocol for optimizing convex functions with polynomial sample complexity exists remained open.

A recent work of Daniely and Feldman [DF18] shows that there exist natural learning problems that are exponentially harder to solve by LDP protocols without interaction. Specifically, they consider PAC learning a class C of Boolean functions over a domain X. A PAC learning algorithm for C receives i.i.d. samples $(x, f^*(x))$ where x is drawn from an unknown distribution D and $f^*\colon X\to\{-1,1\}$, and its goal is to find $\widehat{f}\colon X\to\{-1,1\}$ which achieves a *classification error* of at most α, namely

$$\mathrm{err}_{f^*,D}(\widehat{f}) \doteq \Pr_{x\sim D}[f^*(x)\neq\widehat{f}(x)]\leq\alpha.$$

Daniely and Feldman [DF18] show that the number of samples required by any non-interactive LDP protocol to learn C with a non-trivial error is lower bounded by a polynomial in the margin complexity of C. The margin of a linear separator f over $S\subseteq\mathbb{R}^d$ captures how well the points x with $f(x)=1$ are separated from those with $f(x)=-1$, and is formally defined as

$$\gamma(f,S) \doteq \sup_{w\neq\mathbf{0}}\inf_{x\in S} f(x)\frac{\langle x,w\rangle}{\|x\|_2\|w\|_2}. \tag{1}$$

The margin complexity of C is the inverse of the largest margin γ that can be achieved by embedding X into $\mathbb{R}^d$ such that every $f\in C$ can be realized as a linear separator with margin at least γ. It is a well-studied notion within learning theory and communication complexity, measuring the complexity of Boolean function classes and their corresponding sign matrices in (e.g. [Nov62; ABR64; BGV92; FSS01; BES02; She08; LS09; KS11]). There exist known classes of functions, as decision lists and general linear separators, that are PAC learnable by (interactive) SQ algorithms but have exponentially large margin complexity. Thus, non-interactive LDP protocols require an exponentially larger number of samples for PAC learning such classes than interactive ones. This result also leads to the question of whether all classes with inverse polynomial margin complexity can be learned efficiently non-interactively (see [DF19] for a more detailed discussion). Such large-margin linear classifiers are much more common in practice and are significantly easier to learn than general linear separators. For example, a simple Perceptron algorithm can be used instead of the more involved algorithms like the Ellipsoid method that are used when the margin is exponentially small.

1.1 Our results

We show that both learning large-margin linear separators and learning of linear models with a convex loss require an exponential number of samples in the non-interactive LDP model. Formally, we define the margin relative to a distribution on $\mathbb{R}^d$ as the margin relative to the support of the distribution: $\gamma(f,D)\doteq\gamma(f,\mathrm{supp}(D))$. We give the following lower bound for learning large-margin linear classifiers.

Theorem 1. *Fix $\epsilon>0, \gamma\in(0,1/4]$, $r\in(0,1)$ and $d\geq\gamma^{-2-2r/5}$. Let $\mathcal{A}$ be a randomized, non-interactive ϵ-LDP learning algorithm over $X=\{-1,1\}^d$ using n samples. Assume that for any linear separator f^**

and distribution D over X with margin $\gamma(f^*, D) \geq \gamma$, $\mathcal{A}$ outputs a hypothesis $\widehat{f}$ with an expected error of $\mathbb{E}_{\mathcal{A}}[\mathrm{err}_{f^*,D}(\widehat{f})] \leq 1/2 - \gamma^{1-r}$. Then, $n \geq \exp(C\gamma^{-2r/5})/e^{2\epsilon}$, where $C > 0$ depends only on r.

In particular, this lower bound is always exponential either in the margin or in the dimension of the problem. Note that linear separators with margin γ can be learned with error α by an ϵ-LDP algorithm with $O(1/\gamma^2)$ rounds of interaction and using $\mathrm{poly}(1/(\epsilon\alpha\gamma))$ samples. This can be done by using a standard SQ implementation of the Perceptron algorithm [Blu+97; FGV15] (after a random projection to remove the dependence on the dimension) or via a reduction to convex loss minimization described below together with an LDP algorithm for convex optimization from [DJW13]. Our lower bound is also essentially tight in terms of the achievable error. There exist an efficient non-interactive algorithm achieving an error of $1/2 - \Theta(\gamma)$, while $1/2 - \gamma^{1-r}$ is impossible for all $r > 0$.

Proof technique: As in the prior work [Kas+11; DF18], we exploit the connection to statistical query algorithms. Here, we assume a distribution P over $Z = X \times Y$ and instead of i.i.d. samples from P, an SQ algorithm has access to an SQ oracle for P. Given a query function $h\colon Z \to [-1, 1]$ an SQ oracle for P with tolerance parameter τ returns the value $\mathbb{E}_{z\sim P}[h(z)]$ with some added noise of magnitude bounded by τ [Kea98]. Such an algorithm is non-adaptive if its queries do not depend on the answers to prior queries. Our lower bound is effectively a lower bound against non-adaptive statistical query algorithms together with the known simulation of a non-interactive LDP protocol by a non-adaptive SQ algorithm [Kas+11]. The SQ model captures a broad class of learning algorithms and thus our lower bound can be viewed as showing the importance of interactive access to data beyond the distributed learning setting.

Our lower bound for non-adaptive SQ algorithms is based on a new technique for constructing hard to distinguish pairs of distributions over data. The key technical element of this construction is a pair of distributions over $\{-1, 1\}^d$ that have nearly matching moments but whose supports are nearly linearly separable with significant margin. To design such distributions we rely on tools from the classical moment problem (see Sec. 2.3 for details). A more detailed overview of the proof requires some of the preliminaries and appears in Section 3.1.

Convex loss optimization of linear models: We now spell out the implications of our lower bound in Theorem 1 for stochastic convex optimization. Our lower bounds will apply to optimization of the simple class of *convex linear models*. These models are defined by some loss function $\ell(w, (x, y)) = \varphi(\langle w, x\rangle, y)$ for some φ that is convex in the first parameter for every y. In our reduction the label is in $\{-1, 1\}$ and the loss function can be further simplified as $\ell(w; (x, y)) = \varphi(y\langle w, x\rangle)$ for a fixed convex function $\varphi\colon [-1, 1] \to \mathbb{R}$. In our reduction w and x are in B_d, the unit ball of $\mathbb{R}^d$. We show that there exists L-Lipschitz, σ-smooth and μ-strongly convex φ such that the following lower bound holds.

Theorem 2. *For any parameters $0 \leq \mu < \sigma \leq \infty$, $L > 0$ and $\alpha > 0$, there exists a loss function $\ell(w, (x, y)) = \varphi(y\langle w, x\rangle)$ where φ is convex, L-Lipschitz, σ-smooth and μ-strongly convex, such that any non-interactive ϵ-LDP algorithm $\mathcal{A}$ that outputs $\widehat{w}$ satisfying $\mathbb{E}_{\mathcal{A}}[\ell(\widehat{w}, P)] \leq \inf_{w\in B_d} \mathbb{E}[\ell(w, P)] + \alpha$, requires*

$$n \geq \min\left(\exp\left(cd^{0.16}\right), \exp\left(c\left(\frac{\min(L, \sigma)}{\max(\mu, \alpha)}\right)^{0.19}\right)\right),$$

samples, where $c > 0$ is a universal constant.

This implies that with 1-Lipschitzness and 1-smoothness, the sample complexity is exponential either in d or in $1/\alpha$, and if we add the assumption of μ-strong convexity, the sample complexity can be exponential in $\kappa \doteq \sigma/\mu$. For comparison, for general convex functions the only known upper bounds are exponential in the dimension [STU17; WGX18]. For linear models, by polynomial approximation it is possible to obtain bounds without an exponential dependence in the dimension: for example, Zheng et al. [ZMW17] showed that logistic regression can be solved with roughly $n = \alpha^{-O(\log(1/\alpha))}$ samples and Wang et al. [WSX19] study general

linear models[2]. Efficient non-interactive LDP algorithms exist for least squares linear regression [STU17] and principal component analysis [WX19] since for these tasks low order statistics suffice for finding a solution.

Communication constrained setting: An additional benefit of proving the lower bound via statistical queries is that we can extend our results to other models known to be related to statistical queries. In particular, we consider distributed protocols in which only a small number of bits is communicated from each client. Namely, each client applies a function with range $\{0,1\}^\ell$ to their input and sends the result to the server (for some $\ell \ll \log |Z|$). As the server only has to communicate a random seed which is practically small and can provably be compressed to $O(\log\log|Z| + \log n)$ bits, this model is useful when the communication cost is high and the complete sample $z \in Z$ is expensive to send, for example, when its dimension is large. In the context of learning this model was introduced by Ben-David and Dichterman [BD98] and generalized by Steinhardt et al. [SVW16]. Identical and closely related models are often studied in the context of distributed statistical estimation with communication constraints (e.g. [Luo05; RWV06; RG06; Zha+13; SD15; Sur+16; ACT18; ASZ19; Ach+19; ACT19]). As in the setting of LDP, the number of rounds of interaction that the server uses to solve a learning problem is a critical resource. Using the equivalence between this model and SQ learning we immediately obtain analogous lower bounds for this model. In particular, we show that either $\ell \geq \Omega(\gamma^{-0.39})$ or $n \geq \exp(\Omega(\gamma^{0.39}))$ is required for learning non-interactively. See Section 5 for additional details.

Future work: Our work provides nearly tight lower bounds for learning by non-interactive or one-round LDP protocols. An important question left open is whether linear classification and convex optimization can be solved by algorithms using a small number of rounds of interaction in the above models. Such lower bounds are not known even for the harder problem considered in [DF18]. In contrast, known techniques for solving these problems require a polynomial number of rounds (see [STU17] for a discussion). We hope that the construction in this paper will provide a useful step toward lower bounds against multi-round SQ or LDP algorithms. We remark, however, that general multi-round LDP protocols can be stronger than statistical query algorithms [JMR19] and thus may require an entirely different approach (see discussion in Section 2.1 for more details).

1.2 Related work

Most positive results for non-interactive LDP model concern relatively simple data analysis tasks, such as computing counts and histograms (e.g. [HKR12; EPK14; BS15; BNS18; Erl+18]). Efficient non-interactive algorithms for learning large-margin classifiers and convex linear models can be obtained given access to public unlabeled data [DF18; Wan+19]. A number of lower bounds on the sample complexity of LDP algorithms demonstrate that (non-interactive) LDP protocols are less efficient than the central model of differential privacy [Kas+11; DWJ13; Ull18; DR19].

Joseph et al. [Jos+19; JMR19] explore a different aspect of interactivity in LDP. Specifically, they distinguish between two types of interactive protocols: fully-interactive and sequentially-interactive ones. Fully-interactive protocols place no restrictions on interaction whereas sequentially-interactive ones only allows asking one query per user. They give a separation showing that sequentially-interactive protocols may require exponentially more samples than fully interactive ones. This separation is orthogonal to ours since our lower bounds are against completely non-interactive protocols and we separate them from sequentially-interactive protocols. Acharya et al. [ACT18] implicitly consider another related model: *one-way non-interactive protocols* where the server does not communicate the choice of a randomizer to the clients or, equivalently, cannot share a random string with clients. They give a polynomial separation between one-way non-interactive protocols and non-interactive protocols for the problem of identity testing for a discrete distribution over k elements ($O(k)$ vs $\Omega(k^{3/2})$ samples).

[2]The bound stated by Wang et al. [WSX19] is $n = \alpha^{-O(\log(1/\alpha))}$ for arbitrary 1-Lipschitz losses, contradicting the lower bound in Thm. 2. The authors have confirmed a mistake in their analysis and are working on correcting the bound [SWv].

2 Preliminaries

2.1 Models of computation

Local differential privacy: In the local differential privacy (LDP) model [War65; EGS03; Kas+11] it is assumed that each of n users holds a sample of some dataset $(z_1, \ldots, z_n) \in Z^n$. In the general version of the model the users can communicate with the server arbitrarily. The protocol is said to satisfy (ϵ, δ)-LDP if the algorithm that outputs the transcript[3] of the protocol given the dataset $(z_1, \ldots, z_n)$ satisfies the standard definition of (ϵ, δ)-differential privacy [Dwo+06].

We are interested in the non-interactive (one-round) LDP protocols. Such protocols can equivalently be described as non-interactively accessing the following oracle:

Definition 1. *An ϵ-DP local randomizer $R : Z \to W$ is a randomized algorithm that given an input $z \in Z$, outputs a message $w \in W$, such that $\forall z_1, z_2 \in Z$ and $w \in W$, $\Pr[R(z_1) = w] \leq e^{\epsilon} \Pr[R(z_2) = w]$. For a dataset $S \in Z^n$, an LR_S oracle takes as an input an index i and a local randomizer R and outputs a random value w obtained by applying $R(z_i)$. An algorithm is non-interactive ϵ-LDP if it accesses S only via the LR_S oracle with ϵ-DP local randomizers, each sample is accessed at most once and all of its queries are determined before observing any of the oracle's responses.*

We remark that for non-interactive protocols, querying the same sample multiple times (subject to the entire communication satisfying ϵ-DP) does not affect the model. Also for non-interactive protocols, allowing (ϵ, δ)-differential privacy instead of ϵ-DP does not affect the power of the model [BNS18] (as long as δ is sufficiently small).

Statistical queries: The statistical query model of Kearns [Kea98] is defined by having access to a *statistical query oracle* to the data distribution P instead of i.i.d. samples from P. The oracle is defined as follows:

Definition 2. *Given a domain Z, a* statistical query *is any (measurable) function $h \colon Z \to [-1, 1]$. A* statistical query oracle *$STAT_P(\tau)$ with tolerance τ receives a statistical query h and outputs an arbitrary value v such that $|v - \mathbb{E}_{z \sim Z}[h(z)]| \leq \tau$.*

To solve a learning problem in this model an algorithm has to succeed for any oracle's responses that satisfy the guarantees on the tolerance. In other words, the guarantees of the algorithm should hold in the worst case over the responses of the oracle. A randomized learning algorithm needs to succeed for any SQ oracle whose responses may depend on the all queries asked so far but not on the internal randomness of the learning algorithm.

We say that an SQ algorithm is *non-interactive* (or *non-adaptive*) if all its queries are determined before observing any of the oracle's responses. Kasiviswanathan et al. [Kas+11] show that one can simulate a non-interactive ϵ-LDP algorithm using a non-adaptive SQ algorithm.

Theorem 3 ([Kas+11])**.** *Let $\mathcal{A}$ be an ϵ-LPD algorithm that makes non-interactive queries to LR_S for $S \in Z^n$ drawn i.i.d. from some distribution P. Then for every $\delta > 0$ there is a non-adaptive SQ algorithm $\mathcal{A}_{SQ}$ that in expectation makes $O(n \cdot e^{\epsilon})$ queries to $STAT_P(\tau)$ for $\tau = \Theta(\delta/(e^{2\epsilon} n))$ and whose output distribution has a total variation distance of at most δ from the output distribution of $\mathcal{A}$.*

We remark that this simulation extends to interactive LDP protocols as long as they rely on local randomizers with the sum of privacy parameters used on every point being at most ϵ. Such protocols, first defined in [Kas+11] are referred to as compositional ϵ-LDP. They are known to be exponentially weaker than the general interactive LDP protocols although the separation is known only for rather unnatural problems [JMR19]. The converse of this connection is also known: SQ algorithms can be simulated by ϵ-compositional LDP protocols (and this simulation preserves the number of rounds of interaction) [Kas+11].

[3] The transcript is the set of all messages sent in the protocol.

2.2 Boolean Fourier analysis

Boolean Fourier analysis concerns with the Fourier coefficients of functions of Boolean inputs, $h\colon \{-1,1\}^d \to \mathbb{R}$. Let U_d be the uniform distribution over $\{-1,1\}^d$, and for any $S \subseteq [d]$, define the coefficient

$$\widehat{h}(S) = \mathop{\mathbb{E}}_{x\sim U_d}[h(x)\chi_S(x)], \quad \text{where } \chi_S(x) = \prod_{i\in S} x_i.$$

As $\{\chi_S(x)\}_{S\subseteq[d]}$ is an orthonormal basis of the space of functions $f\colon \{-1,1\}^d \to \mathbb{R}$, h can be decomposed as $h(x) = \sum_{S\subseteq[d]} \widehat{h}(S)\chi_S(x)$. *Plancherel's theorem* states that

$$\mathop{\mathbb{E}}_{x\sim U_d}[h(x)g(x)] = \sum_{S\subseteq[d]} \widehat{h}(S)\widehat{g}(S), \tag{2}$$

and *Parseval's theorem* is the special case where $g = h$. For a distribution D over $\{-1,1\}^d$ we define the Fourier coefficient as the coefficients of the function $x \mapsto \Pr_D[x]/\Pr_{U_d}[x]$, namely,

$$\widehat{D}(S) = \mathop{\mathbb{E}}_{x\sim U_d}\left[\frac{\Pr_D[x]}{\Pr_{U_d}[x]}\chi_S(x)\right] = \mathop{\mathbb{E}}_{x\sim D}\chi_S[x]. \tag{3}$$

Lastly, note that for a distribution D and a function h, it follows from Plancherel's theorem that

$$\mathop{\mathbb{E}}_{x\sim D}[h(x)] = \mathop{\mathbb{E}}_{x\sim U_d}\left[h(x)\frac{\Pr_D[x]}{\Pr_{U_d}[x]}\right] = \sum_{S\subseteq[d]} \widehat{D}(S)\widehat{h}(S). \tag{4}$$

2.3 The classical moment problem

Given a probability distribution P and $k \in \mathbb{N}$, it is natural to try and characterize all distributions that have the same first k moments as P, namely, distributions D with $\mathbb{E}_{x\sim D}[x^i] = \mathbb{E}_{x\sim P}[x^i]$ for all $i \in [k]$. There is a great literature in this topic, e.g. [AK65; KN77] (see [BGGP12] for an application in computer science). The study uses the notion of orthogonal polynomials:

Definition 3. *Let P be a probability distribution over $\mathbb{R}$ with all moments finite. We say that a sequence of polynomials $p_0, p_1, \ldots, p_k, \ldots$ are* orthogonal with respect to P *if the satisfy the following:*

- *For all $m \geq 0$, p_m is of degree m and has a positive leading coefficient.*
- *For all $m, \ell \geq 0$, $\mathbb{E}_{x\sim P}[p_m(x)p_\ell(x)] = \mathbb{1}_{m=\ell}$.*

Denote the above sequence of polynomials as the orthogonal polynomials *with respect to P.*

It is known that there is a unique sequence of orthogonal polynomials with respect to P, hence we call them *the orthogonal polynomials* (w.r.t P). Given the orthogonal polynimials $p_0, p_1, \ldots$, define the function $\rho_k\colon \mathbb{R} \to \mathbb{R}$ as follows:

$$\rho_k(x) = \frac{1}{\sum_{i=0}^k p_i(x)^2}. \tag{5}$$

These functions characterize the amount of mass that can be concentrated on the point x by distributions D that match the first $2k$ moments of P:

Theorem 4 ([AK65], Theorem 2.5.2)**.** *Let P be a distribution with finite moments, fix $k \in \mathbb{N}$ and $x \in \mathbb{R}$ and let ρ_k be defined with respect to P. The following holds:*

- *There exists a distribution D matching the first $2k$ moments of P with $\Pr_D[x] = \rho_k(x)$.*
- *Any distribution D that matches the first $2k$ moments of P satisfies: $\Pr_D[x] \leq \rho_k(x)$.*

3 Proof of Theorem 1

Below we state and prove the lower bound on learning with statistical queries. The lower bounds for LDP protocols stated in Theorem 1 follows directly from the reduction in Theorem 3.

Theorem 5. *Let $r \in (0,1)$, $\gamma \in (0, 2^{-1/(1-r)})$, $n \geq \gamma^{-2-2r/5}$ and define $\eta = \gamma^{1-r}$. Let $\mathcal{A}$ be a non-adaptive statistical query algorithm such that for any linear separator f^* and distribution D over $X = \{-1,1\}^{2d}$ with margin $\gamma(f^*, D) \geq \gamma$, returns a hypothesis $\widehat{f}$ with $\mathbb{E}_{\mathcal{A}}[\mathrm{err}_{f^*,D}(\widehat{f})] \leq 1/2 - \eta$. If $\mathcal{A}$ has access to statistical queries with tolerance $\tau = \exp\left(-c\gamma^{-2r/5}\right)$, then $\mathcal{A}$ requires at least $\exp\left(c\gamma^{-2r/5}\right)$ queries, where $c > 0$ is a constant depending only on r.*

3.1 Outline

We start with a brief sketch of the proof. Let $X = \{-1,1\}^{2d}$ and $Y = \{-1,1\}$. Our proof is based on a construction of two distribution D_0 and D_1 over $\{-1,1\}^{2d} \times \{-1,1\}$ and two linear functions f_0 and f_1 that are hard to distinguish but they almost always disagree on the label y. Specifically, the have the following properties:

- Any $(x,y) \in \mathrm{supp}(D_b)$ satisfies $y = f_b(x)$ for $b \in \{0,1\}$, and additionally, f_0 and f_1 have $\Omega(\gamma)$-classification margin over the supports of D_0 and D_1, respectively.
- D_0 and D_1 have nearly the same Fourier coefficients: for any $S \subseteq [2d]$, $|\widehat{D_0}(x) - \widehat{D_1}(x)|$ is exponentially small.
- $f_0(x) \neq f_1(x)$ for nearly all values of x: $\Pr_{(x,y)\sim D_b}[f_0(x) = f_1(x)] = O(\eta)$, for $b \in \{0,1\}$ where $\eta := \gamma^{1-r}$.

Given these two distributions, we can create a hard family of distributions containing many pairs obtained from the original pair by a simple translation. Any efficient SQ algorithm would find most pairs of distributions impossible to distinguish. That is, the algorithm cannot distinguish which of the two distributions in the pair is the correct one. As a consequence, it will not be able to predict the correct label of x for most values of x.

In the rest of this section we describe how D_0 and D_1 are constructed. The construction involves multiple consecutive steps that we describe below. We start with two distributions P and Q over $\mathbb{R}$ that satisfy:

1. P and Q have matching first $2k = \gamma^{-\Omega(1)}$ moments.
2. $\Pr_{p\sim P}[p \geq \gamma] = 1$ and $\Pr_{q\sim Q}[q \leq -\gamma] \geq 1 - O(\eta)$, where $\eta = \gamma^{1-r}$.

The distribution P is a mixture in which the value γ has weight $1-\eta$ and a scaled and shifted exponential distribution defined on $[\gamma, \infty)$ has weight η. To show that there exists a distribution Q which matches the first $2k$ moments of P and satisfies $\Pr_Q[-\gamma] \geq 1 - O(\eta)$, it suffices to show that $\rho_k(-\gamma) \geq 1 - O(\eta)$, where ρ_k is the function from Eq. (5), which is defined by the orthogonal polynomials of P. We calculate these polynomials as a linear combination of the orthogonal polynomials of the exponential distribution, for which a closed formula is known. We remark that instead of the exponential distribution other distributions can be used to get a similar bound on ρ_k.

Based on P and Q, we create two distributions P_1 and P_{-1} over $\{-1,1\}^d$ which satisfy:

- P_1 and P_{-1} nearly match all Fourier coefficients.
- $\Pr_{x\sim P_1}[\sum_i x_i/d \geq \gamma/2] = 1$ and $\Pr_{x\sim P_{-1}}[\sum_i x_i/d \leq -\gamma/2] \geq 1 - O(\eta)$.

To draw $x \sim P_1$ we first draw $p \sim P$ and then draw each bit of x independently with mean p. Similarly, we draw P_{-1} given Q. The Fourier coefficients of P_1 and P_{-1} correspond to the moments of P and Q, respectively: $\widehat{P_1}(S) = \mathbb{E}_P[p^{|S|}]$ and similarly for P_{-1} and Q. Hence the Fourier coefficients of P_1 and P_{-1} nearly match (note that we've only shown that P and Q match the first $2k$ moments, however, the higher

moments are exponentially small and negligible). The second property of P_1 and P_{-1} follows from the second property of P and Q (except with some small failure probability which we can condition out).

Next, we explain the distributions D_0 and D_1 and the functions f_0 and f_1 that appear in the first paragraph: f_0 is defined as a majority over the first d bits, $f_0(x) = \mathrm{sign}(\sum_{i=1}^{d} x_i)$ and f_1 is a majority over the last d bits, $f_1(x) = \mathrm{sign}(\sum_{i=n+1}^{2d} x_i)$. To draw $(x, y) \sim D_0$, we independently draw $y \sim \mathrm{Unif}(\{-1, 1\})$, $z_1 \sim P_1$ and $z_{-1} \sim P_{-1}$. Then, we set $x = (yz_1, yz_{-1})$. We define D_1 nearly the same way, with the only difference that $x = (yz_{-1}, yz_1)$. From the properties of P_1 and P_{-1}, all properties of D_0 and D_1 presented in the first paragraph are satisfied.

3.2 Proof of Theorem 5

We begin with some notations:

- Given a statistical query h, denote $h(D, f) = \mathbb{E}_{x \sim D}(h(x, f(x)))$.
- We use $C, C', c, c_1, \cdots > 0$ to denote universal constants or constants depending only on r. In the proof we will allow redundant constants depending on r (e.g. the advantage will be $C\eta$ rather than η).
- Let $\mathrm{Unif}(A)$ denote the uniform distribution over a finite set A, let d_{TV} denote the total variation distance of two distributions and let $\mathrm{supp}(P)$ denote the support of a probability distribution P.
- In contrast to the presentation in the intro, we conveniently assume that the distributions D are only over X rather than over $X \times Y$.

The general idea is to split the $2d$ bits of x into two bit-sets, each containing d bits. The value of $f^*(x)$ will be a function of one of these sets, however any efficient non-adaptive algorithm would not be capable of finding the correct subset. Moreover, intuitively speaking, the incorrect subset will almost always *lie* by claiming the wrong value for $f^*(x)$.

We begin with two distributions P_1 and P_{-1} that nearly match all Fourier coefficients, however, $\mathrm{sign}(\sum_{i=1}^{2d} x_i) = 1$ for any $x \in \mathrm{supp}(P_1)$ while $\mathrm{sign}(\sum_i x_i) = -1$ with probability $1 - O(\eta)$ for $x \sim P_{-1}$.

Lemma 1. *There exists two distributions, P_1 and P_{-1} over $\{-1, 1\}^d$, such that the following holds:*

1. *$d_{\mathrm{TV}}(P_1, -P_{-1}) \le C\eta$, where $x \sim -P_{-1}$ is obtained by drawing $x \sim P_{-1}$ and outputting $-x$, and $C > 0$ is a constant depending only on r.*
2. *Any $x \sim \mathrm{supp}(P_1)$ satisfies $\sum_i x_i/d \ge C\gamma$.*
3. *P_1 and P_{-1} are nearly indistinguishable: for any $S \in [d]$, $|\widehat{P_1}(S) - \widehat{P_{-1}}(S)| \le \exp(-c\gamma^{-2r/5})$, where $c > 0$ is a constant depending only on r.*

The proof utilizes results from the classical moment problem, and involves calculating the orthogonal polynomials of some distribution, as will be elaborated in Section 3.3.

Given P_1 and P_{-1}, we construct two pairs of distribution-function (f_0, D_0) and (f_1, D_1) which are hard to distinguish, in a sense that will be clear later. The function f_0 is a majority of the first d coordinates, $f_0(x) = \mathrm{sign}(\sum_{i=1}^{d} x_i)$ and f_1 is a majority of the last d bits, $f_1(x) = \mathrm{sign}(\sum_{i=d+1}^{2d} x_i)$. A random $x \sim D_0$ is drawn by drawing independently $y \in \mathrm{Uniform}(\{-1, 1\})$, $z_1 \sim P_1$, $z_{-1} \sim P_{-1}$ and setting $x = (yz_1, yz_{-1})$. Note that $f_0(x) = y$, where y is the value drawn above. Similarly, $x \sim D_1$ is drawn similarly, with the following distinction: $x = (yz_{-1}, yz_1)$. Here, notice that $f_1(x) = y$.

Since P_1 is nearly distributed as $-P_{-1}$, with high probability over $x \sim D_0$, the majority of the first d coordinates of x is almost always the opposite of the majority of the last last d coordinates (and similarly when $x \sim D_1$). In particular, if one does not know whether the true function f^* equals f_0 or f_1, it is impossible to predict $f^*(x)$ given x with probability significantly greater than a half.

Utilizing the fact that the building blocks of D_0 and D_1, namely P_1 and P_{-1}, nearly match their Fourier coefficients, we can generate a family of hard distributions by simple translations of D_0 and D_1: for any

$a \in \{-1,1\}^{2d}$ define the pairs $(f_{a,0}, D_{a,0})$ and $(f_{a,1}, D_{a,1})$ as follows: $f_{a,0} = \text{sign}(\sum_{i=1}^{d} a_i x_i)$ and $x \sim D_{a,0}$ is obtained by drawing $x' \sim D_0$ and setting $x_i = a_i x'_i$ for $i \in [2d]$. Similarly, $f_{a,1} = \text{sign}(\sum_{i=d+1}^{2d} a_i x_i)$ and $D_{a,1}$ is obtained by drawing $x' \sim D_1$ and setting $x_i = a_i x'_i$. The following are simple properties of the defined distributions, which follow mainly from Lemma 1, and are proved in Section 3.4

Lemma 2. *Fix $a \in \{-1,1\}^{2d}$. Then, $D_{a,0}$ and $D_{a,1}$ satisfy the following properties:*

1. $d_{\text{TV}}(D_{a,0}, D_{a,1}) \leq 2d_{\text{TV}}(P_1, -P_{-1}) \leq C\eta$
2. $\gamma(f_{a,0}, D_{a,0}) \geq C\gamma$ *and* $\gamma(f_{a,1}, D_{a,1}) \geq C\gamma$
3. $\Pr_{x\sim D_{a,0}}[f_{a,0}(x) = f_{a,1}(x)] \leq C\eta$ *and* $\Pr_{x\sim D_{a,1}}[f_{a,0}(x) = f_{a,1}(x)] \leq C\eta$

where $C > 0$ depends only on r (recall that $\eta = \gamma^{1-r}$).

Next, we claim that for any set of $\exp(O(\gamma^{-2r/5}))$ statistical queries and for nearly all values of a, the queries will have nearly the same value for both $(f_{a,0}, D_{a,0})$ and $(f_{a,1}, D_{a,1})$. This follows from the fact that P_1 and P_{-1} have all their Fourier coefficient close to each other.

Lemma 3. *Fix a set of statistical queries $h_1, \ldots, h_k$ for $k \leq \exp(c_1\gamma^{-2r/5})$. Then,*

$$\Pr_{a\in\{-1,1\}^{2d}}\left[\exists i \in [k],\ |h_i(D_{a,0}, f_{a,0}) - h_i(D_{a,1}, f_{a,1})| \geq \exp(-c_2\gamma^{-2r/5})\right] \leq \exp(-c_3\gamma^{-2r/5}),$$

where $c_1, c_2, c_3 > 0$ depend only on r.

The proof will be presented in Section 3.5. Next, we define the exact statistical query setting: define the number of allowed queries k and tolerance τ to ensure that the algorithm cannot distinguish between $(f_{a,0}, D_{a,0})$ and $(f_{a,1}, D_{a,1})$: $k = \exp(c_1\gamma^{-2a/5})$ and $\tau = \exp(-c_2\gamma^{-2a/5})$, for the constants c_1, c_2 from Lemma 3. We define the SQ oracle such that it gives the same answers to $(f_{a,0}, D_{a,0})$ and $(f_{a,1}, D_{a,1})$ for most a: given a statistical query h, it acts as follows:

- If the true distribution-function pair is $(f_{a,0}, D_{a,0})$ for some $a \in \{-1,1\}^{2d}$ then return the true value $h(f_{a,0}, D_{a,0})$.
- If the pair is $(f_{a,1}, D_{a,1})$ and $|h(f_{a,0}, D_{a,0}) - h(f_{a,1}, D_{a,1})| \leq \tau$ then return $h(f_{a,0}, D_{a,0})$.
- Otherwise return $h(f_{a,1}, D_{a,1})$.

To conclude the proof, recall that Lemma 2 states that for nearly all values of x, $f_{a,0}(x) \neq f_{a,1}(x)$. In particular, if one cannot distinguish between these two functions, then they cannot know the true classification of x. There some delicacy that should be taken care of: if the total variation distance between $D_{a,0}$ and $D_{a,1}$ was large, it would have been possible, given x, to guess whether it was drawn from $D_{a,0}$ or $D_{a,1}$ with a non-negligible success probability. However, Lemma 2 ensures that this is not the case. The formal proof is presented below:

Proof of Theorem 5. We start by assuming that the algorithm is deterministic and then extend to randomized algorithms. From this assumption it follows that the statistical queries $h_1, \ldots, h_k$ are deterministic as well. Fix a such that the responses of the oracle to $(D_{a,0}, D_{a,1})$ are the same as for $(D_{a,1}, f_{a,1})$. From Lemma 3 and from the definition of the oracle, nearly all a are such. For these a, the algorithm has to learn some hypothesis without knowing if the true distribution-function pair is $(D_{a,0}, f_{a,0})$ or $(D_{a,1}, f_{a,1})$. Let $\mathcal{A}_{a,b}$

denote the learned hypothesis given $(f_{a,b}, D_{a,b})$. For these hard values of a, $\mathcal{A}_{a,0} = \mathcal{A}_{a,1}$. Let $\eta' := C\eta$, where C is the constant from Lemma 2. Applying Lemma 2 multiple times, we obtain that for any such a:

$$\begin{aligned}
&\Pr_{x\sim D_{a,0}}[\mathcal{A}_{a,0}(x) = f_{a,0}(x)] + \Pr_{x\sim D_{a,1}}[\mathcal{A}_{a,1}(x) = f_{a,1}(x)] \\
&\quad= \Pr_{x\sim D_{a,0}}[\mathcal{A}_{a,0}(x) = f_{a,0}(x)] + \Pr_{x\sim D_{a,1}}[\mathcal{A}_{a,0}(x) = f_{a,1}(x)] \\
&\quad\le \Pr_{x\sim D_{a,0}}[\mathcal{A}_{a,0}(x) = f_{a,0}(x)] + \Pr_{x\sim D_{a,1}}[\mathcal{A}_{a,0}(x) \neq f_{a,0}(x)] + \Pr_{x\sim D_{a,1}}[f_{a,1} = f_{a,0}(x)] \\
&\quad\le \Pr_{x\sim D_{a,0}}[\mathcal{A}_{a,0}(x) = f_{a,0}(x)] + \Pr_{x\sim D_{a,1}}[\mathcal{A}_{a,0}(x) \neq f_{a,0}(x)] + \eta' \\
&\quad\le \Pr_{x\sim D_{a,0}}[\mathcal{A}_{a,0}(x) = f_{a,0}(x)] + \Pr_{x\sim D_{a,0}}[\mathcal{A}_{a,0}(x) \neq f_{a,0}(x)] + 2\eta' \qquad (6) \\
&\quad= 1 + 2\eta'.
\end{aligned}$$

where Eq. (6) follows from the fact that $d_{\mathrm{TV}}(D_{a,0}, D_{a,1}) \le \eta'$. From Lemma 3, the above holds for a $1 - \exp(-c\gamma^{2r/5})$-fraction of the values of a (where $c > 0$ depends only on r). In particular,

$$\Pr_{a\sim \mathrm{Unif}(\{-1,1\}^{2d}), b\sim\mathrm{Unif}(\{0,1\}), x\sim D_{a,b}}[\mathcal{A}_{a,b}(x) = f_{a,b}(x)] \le 1/2 + C'\eta', \qquad (7)$$

where C' depends only on r. Lastly, assume that the algorithm is randomized. Any randomized algorithm is just a distribution over deterministic algorithms, hence Eq. (7) will hold even if the algorithm is allowed to be randomized and the probability is taken over a, b, x and the randomness of the algorithm. □

3.3 Proof of Lemma 1

Notation. Throughout the proof we will use the following parameters: $\eta = \gamma^{1-r}$, $\gamma' = \gamma^{1-2r/5}$ and $k = \lfloor(\eta/\gamma')^{2/3}\rfloor = \lfloor\gamma^{-2r/5}\rfloor$. From the assumptions in Theorem 5, $\gamma, \eta, \gamma' \in (0, 1/2]$.

Outline. The first step is to find two distributions over $\mathbb{R}$ of a particular shape that their first k moments match. The first distribution P is a mixture that samples 0 with probability $1-\eta$ and an exponential random variable with probability η. By calculating the orthogonal polynomials of P and applying Theorem 4, we find a distribution Q that matches the first k moments of P, and additionally, $\Pr_Q[-\gamma'] \ge 1 - O(\eta)$.

In the second step, we shift, scale and condition P and Q, to obtain two distributions P' and Q' that have nearly matching moments and satisfy the following conditions: $\Pr_{P'}[\gamma] \ge 1 - O(\eta)$; $\Pr_{Q'}[-\gamma] \ge 1 - O(\eta)$; P' is supported on $[\gamma, 1/2]$ and Q' is supported on $[-1/2, 1/2]$.

In the third step, we use P' and Q' to generate P_1 and P_{-1}, respectively. To generate $x = (x_1, \ldots, x_d) \sim P_{-1}$, we first draw $p \sim Q'$ and then, conditioned on p, we draw each x_i i.i.d. from the distribution over $\{-1, 1\}$ with expectation p. The distribution P_1 is similarly defined using P', except that we additionally condition on the high-probability event that $\sum_i x_i/d \ge \gamma/2$. It follows from a simple argument that the Fourier coefficients satisfy $\widehat{P_{-1}}(S) = \mathbb{E}_{p\sim Q'}[p^{|S|}]$ and similarly, $\widehat{P_1}(S) \approx \mathbb{E}_{p\sim P'}[p^{|S|}]$. We obtain that all Fourier coefficients of P_1 and P_{-1} nearly match.

Lastly, we claim that $d_{\mathrm{TV}}(P_1, -P_{-1}) \le O(\eta)$. To obtain this, first note that $d_{\mathrm{TV}}(P', -Q') \le O(\eta)$, as both P' and $-Q'$ have $1 - O(\eta)$ mass on γ. As P_1 and P_{-1} are obtained from P' and Q' using nearly the same transformation, we can apply the *data processing inequality* (presented in Section A) to bound $d_{\mathrm{TV}}(P_1, P_{-1}) \lesssim d_{\mathrm{TV}}(P', Q')$.

We divide the proof into four parts, according to the steps described above.

Step 1: Distributions P and Q over $\mathbb{R}$ that match the first moments

We start by constructing two distributions over $\mathbb{R}$ with matching first $2k$ moments. Let distribution P be the following mixture: with probability η sample from the exponential distribution with parameter 1, and with probability $1 - \eta$ sample 0. We start with the following lemma:

Lemma 4. *There exists a distribution Q that matches the first $2k$ moments of P and additionally, $\Pr_{x\sim Q}[x = -\gamma'] \geq 1 - C\eta$, where $C > 0$ is a universal constant.*

Before proving this lemma, we give some intuition: By Theorem 4, it suffices to show that $\rho_k(-\gamma') \geq 1 - O(\eta)$, where ρ_k is as defined in Section 2.3 with respect to the moments of P. The same theorem implies that since $\Pr_P[0] = 1-\eta$, then $\rho_k(0) \geq 1-\eta$; and since ρ_k is continuous, $\rho_k(-y) \geq 1-O(\eta)$ for any sufficiently small y. To show that $\rho_k(\gamma') \geq 1-O(\eta)$, we calculate the orthogonal polynomials of P as linear combinations of the Laguerre polynomials, the orthogonal polynomials for the exponential distribution. Recall that ρ_k is defined as a function of these polynomials, which allows us to bound ρ_k.

First, we present the orthogonal polynomials of the exponential distribution:

Lemma 5. *The orthogonal polynomials for the exponential distribution with parameter 1 are the Laguerre polynomials*

$$L_m(x) = \sum_{i=0}^{m} \binom{m}{i} \frac{(-1)^i}{i!} x^i.$$

Using a simple calculation, one obtains that the orthogonal polynomials $\{p_m\}_{m=0}^{\infty}$ for P equal

$$p_m(x) = \mu \left((m + \eta/(1-\eta)) L_m - \sum_{\ell=0}^{m-1} L_\ell(x) \right),$$

where

$$\mu^{-2} = \eta(m + \eta/(1-\eta))^2 + \eta m + \eta^2/(1-\eta) = \eta(m^2+m) + O(m\eta^2).$$

is a normalizing constant. To verify this formula it suffices to check that $\mathbb{E}_P[p_m(x)L_\ell(x)] = 0$ for $\ell < m$ and that $\mathbb{E}_P[p_m(x)^2] = 1$ and these equations uniquely define p_m (up to sign changes).

To get a closed form equation of the orthogonal polynomial, we use the identity

$$\sum_{\ell=i}^{m-1} \binom{\ell}{i} = \binom{m}{i+1},$$

to obtain that

$$\sum_{\ell=0}^{m-1} L_\ell(x) = \sum_{\ell=0}^{m-1} \sum_{i=0}^{\ell} \binom{\ell}{i} \frac{(-1)^i}{i!} x^i = \sum_{i=0}^{m-1} \sum_{\ell=i}^{m-1} \binom{\ell}{i} \frac{(-1)^i}{i!} x^i = \sum_{i=0}^{m-1} \binom{m}{i+1} \frac{(-1)^i}{i!} x^i,$$

hence

$$p_m(x)/\mu = \sum_{i=0}^{m} \left((m + \eta/(1-\eta)) \binom{m}{i} - \binom{m}{i+1} \right) \frac{(-1)^i}{i!} x^i, \tag{8}$$

where $\binom{m}{m+1} = 0$.

Using the above formula, we can prove the following bound on $\rho_k(x)$:

Lemma 6. *Assume that $|x| \leq \eta k^{-3/2}$. Then, $\rho_k(x) \geq 1 - C\eta$, where $C > 0$ is a universal constant.*

Proof. We start by bounding the coefficients of $p_m(x)$. Denote $p_m(x) = \sum_{i=0}^{m} \xi_{m,i} x_i$. For any $i \leq m$, we use that fact that

$$\binom{m}{i+1} = \frac{m-i}{i+1} \binom{m}{i} \leq (m + \eta/(1-\eta)) \binom{m}{i}$$

and Eq. (8) to estimate

$$|\xi_{m,i}| \leq \mu(m + \eta/(1-\eta)) \binom{m}{i} / i!.$$

Additionally, it follows from definition of μ that

$$\mu \leq \frac{1}{\sqrt{\eta}(m + \eta/(1-\eta))}, \tag{9}$$

hence

$$|\xi_{m,i}| \leq \binom{m}{i} \frac{1}{i!\sqrt{\eta}}.$$

For $i = 0$ and $m > 0$ we can get a tighter bound using more accurate calculation, Eq. (9) and $\eta \leq 1/2$:

$$|\xi_{m,0}| = \frac{\mu\eta}{1-\eta} \leq \frac{\sqrt{\eta}}{m(1-\eta)} \leq \frac{2\sqrt{\eta}}{m}.$$

We proceed with bounding p_m for $m > 0$, using the inequality $\binom{m}{i} \leq m^i/i!$:

$$|p_m(x)| \leq \sum_{i=0}^{m} |\xi_{m,i}| x^i \leq \frac{2\sqrt{\eta}}{m} + \sum_{i=1}^{m} \binom{m}{i} \frac{x^i}{i!\sqrt{\eta}} \leq \frac{2\sqrt{\eta}}{m} + \sum_{i=1}^{m} \frac{(mx)^i}{(i!)^2\sqrt{\eta}}.$$

For any $1 \leq m \leq k$, by the requirement of this lemma, $mx \leq kx \leq \eta/\sqrt{k} \leq 1$, hence

$$|p_m(x)| \leq \frac{2\sqrt{\eta}}{m} + \sum_{i=1}^{\infty} \frac{mx}{(i!)^2\sqrt{\eta}} \leq \frac{2\sqrt{\eta}}{m} + \frac{Cmx}{\sqrt{\eta}} \leq \frac{2\sqrt{\eta}}{m} + C\sqrt{\frac{\eta}{k}},$$

where $C > 0$ is a universal constant. Using the inequality $(a+b)^2 \leq 2a^2 + 2b^2$ we obtain that

$$\sum_{m=1}^{k} p_m(x)^2 \leq \sum_{m=1}^{k} \left(\frac{8\eta}{m^2} + \frac{2C^2\eta}{k} \right) = C'\eta,$$

for a universal $C' > 0$. Since $p_0 \equiv 1$, we get that

$$\rho_k(x) = \frac{1}{\sum_{m=0}^{k} p_m(x)^2} \geq \frac{1}{1 + C'\eta} \geq 1 - C'\eta,$$

as required. □

By definition, $\gamma' \leq \eta/k^{3/2}$. Combining Lemma 6 and Theorem 4, the proof of Lemma 4 concludes.

Step 2: Re-scaling and restricting P and Q to obtain P' and Q'

For any $\alpha, \beta \in \mathbb{R}$, let $\alpha P + \beta$ denote the distribution obtained in the obvious manner, by drawing $x \sim P$ and outputting $\alpha x + \beta$. In the same fashion, let P' denote the distribution $(P + \gamma'/2)/(8k+1)$ conditioned on $[-1/2, 1/2]$ and Q' denote $(Q + \gamma'/2)/(8k+1)$ conditioned on $[-1/2, 1/2]$. Let $\widetilde{\gamma} = \gamma'/(16k+2)$ and note that $\widetilde{\gamma} = \Theta(\gamma)$. The following holds with respect to P' and Q':

Lemma 7. *The following holds:*

- $P'(\widetilde{\gamma}) \geq 1 - C\eta$ *and* $Q'(-\widetilde{\gamma}) \geq 1 - C\eta$.
- *For any integer* $i \geq 0$, $|E_{P'}[x^i] - \mathbb{E}_{Q'}[x^i]| \leq e^{-ck}$.

(where $c, C > 0$ are universal constants.)

Before proceeding with the proof, here is an intuition: the first item follows from the definitions of P, P' and Q' and Lemma 4. For the second second item, note that the first k moments of P' and Q' nearly match because P and Q match these moments, and the remaining moments nearly match since they are small, since P' and Q' are supported on $[-1/2, 1/2]$. In the proof we argue that conditioning on $[-1/2, 1/2]$ does not matter much, by obtaining tail bounds on P and Q, using a generalized Markov's inequality based on their first $2k$ moments.

Proof of Lemma 7. The first item follows from $P(0) = 1 - \eta$, Lemma 4 which states that $Q(-\gamma') \geq 1 - C\eta$, and from the definitions of P', Q' and $\widetilde{\gamma}$.

Next, we prove the second item of the lemma. First, it is an easy exercise to check that $(P+\gamma'/2)/(8k+1)$ and $(Q+\gamma'/2)/(8k+1)$ match the first k moments, as P and Q do. Next, we argue that conditioning on $[-1/2, 1/2]$ does not change the first k moments considerably, which would imply that P' and Q' nearly match those moments. This is obtained by bounding the tails of $(P+\gamma'/2)/(8k+1)$ and $(Q+\gamma'/2)/(8k+1)$. For that purpose, note that moment m of the exponential distribution equals $m!$, hence,

$$\mathop{\mathbb{E}}_{P}[x^{2k}] = \eta(2k)! \leq (2k)^{2k}. \tag{10}$$

Using $\gamma' \leq 1$, Markov's inequality, the fact that P and Q match the first $2k$ moments and Eq. (10), we obtain that for any $t \geq 1/2$,

$$\begin{aligned}\Pr_{(Q+\gamma'/2)/(8k+1)}[|x| \geq t] &\leq \Pr_{Q}[|x| \geq t(8k+1) - \gamma'/2] \leq \Pr_{Q}[|x| \geq 8kt] \\ &= \Pr_{Q}\left[x^{2k} \geq (8kt)^{2k}\right] \leq \mathop{\mathbb{E}}_{Q}\left[x^{2k}\right]/(8kt)^{2k} = \mathop{\mathbb{E}}_{P}\left[x^{2k}\right]/(8kt)^{2k} \leq (4t)^{-2k}.\end{aligned} \tag{11}$$

Using Eq. (11) it is simple to see that for any moment $m \in [k]$,

$$\left|\mathop{\mathbb{E}}_{(Q+\gamma'/2)/(8k+1)}[x^m] - E_{Q'}[x^m]\right| \leq e^{-ck} \tag{12}$$

for some universal constant $c > 0$. From definition of P, it is also easy to see that

$$\left|\mathop{\mathbb{E}}_{(P+\gamma'/2)/(8k+1)}[x^m] - E_{P'}[x^m]\right| \leq e^{-ck}. \tag{13}$$

Eq. (12), Eq. (13) and the fact that $(Q+\gamma'/2)/(8k+1)$ and $(P+\gamma'/2)/(8k+1)$ match their first $2k$ moments, imply that $|\mathbb{E}_{Q'}[x^m] - \mathbb{E}_{P'}[x^m]| \leq 2e^{-ck}$ for any $m \in [k]$.

Lastly, it remains to argue that P' and Q' nearly match the moments $m > k$. Indeed,

$$|\mathop{\mathbb{E}}_{Q'}[x^m] - \mathop{\mathbb{E}}_{P'}[x^m]| \leq |\mathop{\mathbb{E}}_{Q'}[x^m]| + |\mathop{\mathbb{E}}_{P'}[x^m]| \leq 2^{-m} + 2^{-m} \leq 2^{-k+1},$$

using the fact that P' and Q' are supported on $[-1/2, 1/2]$. □

Step 3: moving from $\mathbb{R}$ to the Boolean cube

Using the distributions P' and Q' we define distributions P_B, Q_B over the boolean cube $\{-1,1\}^d$, where $x \sim P_B$ is drawn as follows: first, we draw $p \sim P'$. Conditioned on p, each bit x_i is drawn independently such that $\mathbb{E}[x_i \mid p] = p$. Equivalently, $\Pr[x_i = 1] - \Pr[x_i = -1] = p$. Similarly, Q_B is defined when P' is replaced with Q'. We obtain that for any set $S \subseteq [d]$,

$$\widehat{P_B}(S) = \mathop{\mathbb{E}}_{x\sim P_B}\left[\prod_{i\in S} x_i\right] = \mathop{\mathbb{E}}_{p\sim P}\left[\mathbb{E}\left[\prod_{i\in S} x_i \;\middle|\; p\right]\right] = \mathop{\mathbb{E}}_{p\sim P}\left[\prod_{i\in S}\mathbb{E}\left[x_i \mid p\right]\right] = \mathop{\mathbb{E}}_{p\sim P}\left[p^{|S|}\right],$$

and similarly, $\widehat{Q_B}(S) = \mathbb{E}_Q[p^{|S|}]$. Hence, Lemma 7 implies that $|\widehat{P_B}(S) - \widehat{Q_B}(S)| \leq e^{-ck}$ for any $S \subseteq [d]$.

Notice that P_B and Q_B almost satisfy the requirements of Lemma 1 as P_1 and P_{-1}, however, it is required that $\mathrm{sign}(\sum_i x_i) \geq \Omega(\gamma)$ for any $x \in \mathrm{supp}(P_1)$. Hence, we define the distribution P'_B which equals P_B conditioned on $\sum_i x_i/d \geq \widetilde{\gamma}/2 = \Omega(\gamma)$. The conditioning does not change the distribution considerably: since $p \sim P'$ always satisfies $p \geq \widetilde{\gamma}$, we obtain by Chernoff's inequality that

$$\Pr_{x\sim P_B}\left[\frac{1}{d}\sum_i x_i \leq \widetilde{\gamma}/2\right] = \mathop{\mathbb{E}}_{p\sim P}\left[\Pr\left[\frac{1}{d}\sum_i x_i \leq \widetilde{\gamma}/2 \;\middle|\; p\right]\right] \leq e^{-d\widetilde{\gamma}^2/8} \leq e^{-kc}, \tag{14}$$

for a universal constant $c > 0$, using the assumption $d \geq \Omega(k/\gamma^2) \geq \Omega(k/\widetilde{\gamma}^2)$. In particular, $|\widehat{P_B}(S) - \widehat{P'_B}(S)| \leq e^{-ck}$ for any S, which implies, by the triangle inequality, that

$$|\widehat{Q_B}(S) - \widehat{P'_B}(S)| \leq |\widehat{Q_B}(S) - \widehat{P_B}(S)| + |\widehat{P_B}(S) - \widehat{P'_B}(S)| \leq e^{-ck}$$

for some other $c > 0$. We set $P_1 = P'_B$ and $P_{-1} = Q_B$, and we have shown that these distributions satisfy statements 2 and 3 of Lemma 1.

Step 4: bounding the total variation between P_1 and P_{-1}

It remains to bound the total variation between P_1 and $-P_{-1}$. We will use the *data processing inequality*, presented in Section A. By that inequality, as P_B and Q_B are obtained from P' and Q' using the same transformation, we obtain that

$$d_{\mathrm{TV}}(P_B, -Q_B) \leq d_{\mathrm{TV}}(P', -Q') \leq C\eta$$

for some universal $C > 0$, where the last inequality follows from the fact that P' and $-Q'$ both have $1 - O(\eta)$ mass on $\widetilde{\gamma}$. From Eq. (14) and the definition of P'_B it follows that $d_{\mathrm{TV}}(P_B, P'_B) \leq e^{-ck}$, hence we get by the triangle inequality that

$$d_{\mathrm{TV}}(P_1, P_{-1}) = d_{\mathrm{TV}}(P'_B, P_Q) \leq d_{\mathrm{TV}}(P'_B, P_B) + d_{\mathrm{TV}}(P_B, P_Q) \leq C\eta + e^{-ck} \leq C'(r),$$

where $C'(r)$ is a constant that depends only on r, and the last inequality follows from the fact that k and $1/\eta$ are polynomially related for a fixed r, hence $e^{-ck} \leq C'(r)\eta$.

3.4 Proof of Lemma 2

We start with proving statement 1 of the lemma. First, for any four probability distributions P, P', Q, Q' defined on the same probability space, $d_{\mathrm{TV}}(P \times Q, P' \times Q') \leq d_{\mathrm{TV}}(P, P') + d_{\mathrm{TV}}(Q, Q')$, hence $d_{\mathrm{TV}}(P_1 \times P_{-1}, (-P_{-1}) \times (-P_1)) \leq 2d_{\mathrm{TV}}(P_1, -P_{-1})$. Next, note that D_0 is obtained by drawing $x \sim P_1 \times P_{-1}$, drawing $y \sim \mathrm{Unif}(\{-1, 1\})$ and outputting yx, and D_1 is obtained from $(-P_{-1}) \times (-P_1)$ the same way. Hence, from the data processing inequality (Lemma 11 in Section A),

$$d_{\mathrm{TV}}(D_0, D_1) \leq d_{\mathrm{TV}}(P_1 \times P_{-1}, (-P_{-1}) \times (-P_1)) \leq 2d_{\mathrm{TV}}(P_1, -P_{-1}).$$

Next, note that $D_{a,0}$ is obtained from D_0 the same way that $D_{a,1}$ is obtained from D_1, hence, by the data processing inequality,

$$d_{\mathrm{TV}}(D_{a,0}, D_{a,1}) \leq d_{\mathrm{TV}}(D_0, D_1) \leq 2d_{\mathrm{TV}}(P_1, -P_{-1}) \leq C\eta.$$

For statement 2, let $a' = (a_1, \ldots, a_d, 0, \ldots, 0)$. Then,

$$\begin{aligned}\gamma(f_{a,0}, D_{a,0}) &\geq \inf_{x \in \mathrm{supp}(D_{a,0})} f_{a,0}(x) \frac{x^\top a'}{\|x\|\|a'\|} = \inf_{x \in \mathrm{supp}(D_0)} f_{a,0}(x) \frac{(x_1 a_1, \ldots, x_{2d} a_{2d})^\top a'}{\|x\|\|a'\|} \\ &= \inf_{x \in \mathrm{supp}(D_0)} f_{a,0}(x) \frac{\sum_{i=1}^d x_i}{\|x\|\|a'\|} = \inf_{x \in \mathrm{supp}(P_1)} \frac{\left|\sum_{i=1}^d x_i\right|}{\|x\|\|a'\|} = \inf_{x \in \mathrm{supp}(P_1)} \frac{\left|\sum_{i=1}^d x_i\right|}{\sqrt{2}d} \geq \frac{C\gamma}{\sqrt{2}},\end{aligned}$$

where we used Lemma 1 for the last inequality. Similarly, we can lower bound $\gamma(f_{a,1}, D_{a,1})$, using $a' = (0, \ldots, 0, a_{d+1}, \ldots, a_{2d})$.

Lastly, we prove statement 3. To simplify notation, we will assume that $a = \mathbf{1}$ (the all-ones vector), however, it is simple to see that the statement holds for any a. Recall that $x \sim D_{\mathbf{1},0}$ is drawn by drawing $z_1 \sim P_1$, $z_{-1} \sim P_{-1}$, $y \sim \mathrm{Unif}(\{-1, 1\})$ and setting $x = (yz_1, yz_{-1})$. From section 2 of Lemma 1,

$$f_{\mathbf{1},0}(x) = \mathrm{sign}\left(y \sum_{i=1}^d (z_1)_i\right) = y. \tag{15}$$

From section 1 or Lemma 1, $d_{\mathrm{TV}}(P_1, -P_{-1}) \le C\eta$, which implies that with probability $1-C\eta$, $\mathrm{sign}(\sum_{i=1}^{d}(z_{-1})_i) = -1$, hence, with probability $1 - C\eta$,

$$f_{\mathbf{1},1}(x) = \mathrm{sign}\left(y\sum_{i=1}^{d}(z_{-1})_i\right) = -y. \tag{16}$$

From Eq. (15) and Eq. (16), the proof follows. A similar statement holds when we replace $D_{\mathbf{1},0}$ with $D_{\mathbf{1},1}$, $D_{a,0}$ and $D_{a,1}$.

3.5 Proof of Lemma 3

We will be considering one statistical query.

Lemma 8. *Let $\theta := 2\max_{S\subseteq[d]}|\widehat{P}_1(S) - \widehat{P}_{-1}(S)|$. Fix a statistical query $h\colon \{-1,1\}^{2d}\times\{-1,1\}\to[-1,1]$. Then, for any $t>0$,*

$$\Pr_{a\sim\mathrm{Unif}(\{-1,1\}^{2d})}[|h(D_{a,0}, f_{a,0}) - h(D_{a,1}, f_{a,1})| \ge t] \le 2\theta^2/t^2.$$

Define the conditional distribution $D_{a,b|y=1}$ and as the conditional distribution of $x\sim D_{a,b}$ given $f_{a,b}(x) = 1$ and similarly define $D_{a,b|y=-1}$. This enables us to decompose any statistical query h in two: $h_1(x) = h(x,1)$ and $h_{-1}(x) = h(x,-1)$. Note that

$$\begin{aligned} h(D_{a,b}, f_{a,b}) &= \mathop{\mathbb{E}}_{x\sim D_{a,b}}[h(x, f_{a,b}(x))] \\ &= \Pr_{D_{a,b}}[f_{a,b}(x)=1]\mathop{\mathbb{E}}_{D_{a,b|y=1}}[h_1(x)] + \Pr_{D_{a,b}}[f_{a,b}(x)=-1]\mathop{\mathbb{E}}_{D_{a,b|y=-1}}[h_{-1}(x)] \\ &= \frac{1}{2}h_1(D_{a,b|y=1}) + \frac{1}{2}h_{-1}(D_{a,b|y=-1}). \end{aligned} \tag{17}$$

Next, we present and prove a simple lemma:

Lemma 9. *Let P, P', Q and Q' be distributions over $\{-1,1\}^d$. Then, for any $S_1, S_2\subseteq[d]$,*

$$|\widehat{P\times Q}(S_1,S_2) - \widehat{P'\times Q'}(S_1,S_2)| \le |\widehat{P}(S_1) - \widehat{Q}(S_1)| + |\widehat{P}(S_2) - \widehat{Q}(S_2)|.$$

Proof. Note that

$$\widehat{P\times Q}(S_1,S_2) = \mathop{\mathbb{E}}_{x_1\sim P, x_2\sim Q}[\chi_{S_1,S_2}(x_1,x_2)] = \mathop{\mathbb{E}}_{x_1\sim P}[\chi_{S_1}(x_1)]\mathop{\mathbb{E}}_{x_2\sim Q}[\chi_{S_2}(x_2)] = \widehat{P}(S_1)\widehat{Q}(S_2).$$

Hence,

$$\begin{aligned} |\widehat{P\times Q}(S_1,S_2) - \widehat{P'\times Q'}(S_1,S_2)| &= |\widehat{P}(S_1)\widehat{Q}(S_2) - \widehat{P'}(S_1)\widehat{Q'}(S_2)| \\ &\le |\widehat{P}(S_1)||\widehat{Q}(S_2) - \widehat{Q'}(S_2)| + |\widehat{Q'}(S_2)||\widehat{P}(S_1) - \widehat{P'}(S_1)|. \end{aligned}$$

From Eq. (3), each Fourier coefficient of a probability distribution is bounded by 1 in absolute value, and the proof follows. □

As $|\widehat{P_1}(S) - \widehat{P_{-1}}(S)| \le \theta/2$ for all $S\subseteq[d]$, we obtain from Lemma 9 that:

$$\forall S\subseteq[2d]\colon\ |\widehat{D_{\mathbf{1},0|y=1}}(S) - \widehat{D_{\mathbf{1},1|y=1}}(S)| = |\widehat{P_1\times P_{-1}}(S) - \widehat{P_{-1}\times P_1}(S)| \le \theta, \tag{18}$$

where $\mathbf{1}$ is the all-ones vector. We use this inequality to prove the following lemma:

Lemma 10. *Fix $h_1\colon\{-1,1\}^{2d}\to[-1,1]$ and $t>0$. Then,*

$$\Pr_{a\sim\mathrm{Unif}(\{-1,1\}^{2d})}[|h_1(D_{a,0|y=1}) - h_1(D_{a,1|y=1})| \ge t] \le \theta^2/t^2.$$

Proof. Denote $P = D_{\mathbf{1},0|y=1}$ and $Q = D_{\mathbf{1},1|y=1}$. Denote $P_a = D_{a,0|y=1}$ and $Q_a = D_{a,1|y=1}$, and notice that $x \sim P_a$ is obtained by drawing $x' \sim P$ and setting $x_i = a_i x'_i$, and similarly Q_a is obtained from Q. Hence $\widehat{P_a}(S) = \widehat{P}(S)\chi_S(a)$, where $\chi_S(a) = \prod_{i\in S} a_i$. Similarly, $\widehat{Q_a}(S) = \widehat{Q}(S)\chi_S(a)$.

From Eq. (4),

$$h_1(P_a) - h_1(Q_a) = \sum_S (\widehat{P}_a(S) - \widehat{Q}_a(S))\widehat{h_1}(S) = \sum_S (\widehat{P}(S) - \widehat{Q}(S))\widehat{h_1}(S)\chi_S(a).$$

Squaring both sides, taking expectation over a, we obtain that

$$\begin{aligned}
\mathop{\mathbb{E}}_a (h_1(P_a) - h_1(Q_a))^2 &= \mathop{\mathbb{E}}_a \left(\sum_S \widehat{h_1}(S)(\widehat{P}_a(S) - \widehat{Q}_a(S)) \right)^2 \\
&= \sum_{S,T} \widehat{h_1}(S)\widehat{h_1}(T)(\widehat{P}(S) - \widehat{Q}(S))(\widehat{P}(T) - \widehat{Q}(T)) \mathop{\mathbb{E}}_a \chi_S(a)\chi_T(a) \\
&= \sum_S \widehat{h_1}(S)^2(\widehat{P}(S) - \widehat{Q}(S))^2 && (19) \\
&\le \theta^2 \sum_S \widehat{h_1}(S)^2 && (20) \\
&= \theta^2 \mathop{\mathbb{E}}_{x\in\{-1,1\}^{2d}} [h(x)^2] && (21) \\
&\le \theta^2. && (22)
\end{aligned}$$

where Eq. (19) follows from $\mathbb{E}_a \chi_S(a)\chi_T(a) = \mathbb{1}_{S=T}$, Eq. (20) follows from Eq. (18) and the definitions of P and Q, Eq. (21) is Parseval's equality (Eq. (2)) and Eq. (22) is due to the fact that by definition, $h_1(x) \in [-1, 1]$ for all x. Therefore, by Chebyshev's inequality,

$$\Pr_a[h_1(D_{a,0|y=1}) - h_1(D_{a,1|y=1})| \ge t] = \Pr_a[|h_1(P_a) - h_1(Q_a)| > t] \le \mathrm{Var}(h_1(P_a) - h_1(Q_a))/t^2 \le \theta^2/t^2.$$

□

We are ready to conclude the proof of Lemma 8:

Proof of Lemma 8. From Eq. (17),

$$\begin{aligned}
&\Pr_a[|h(D_{a,0}, f_{a,0}) - h(D_{a,1}, f_{a,1})| \ge t] \\
&\quad\le \Pr_a[|h_1(D_{a,0|y=1}) - h_1(D_{a,1|y=1})| \ge t] + \Pr_a[|h_{-1}(D_{a,0|y=-1}) - h_{-1}(D_{a,1|y=-1})| \ge t] \\
&\quad= \Pr_a[|h_1(D_{a,0|y=1}) - h_1(D_{a,1|y=1})| \ge t] + \Pr_a[|h_{-1}(D_{-a,0|y=1}) - h_{-1}(D_{-a,1|y=1})| \ge t] && (23) \\
&\quad\le 2\theta^2/t^2, && (24)
\end{aligned}$$

where Eq. (23) follows from $h_{-1}(D_{a,b|y=1}) = h_{-1}(D_{-a,b}|y = 1)$ and Eq. (24) follows from Lemma 10. □

Lastly, we conclude the proof of Lemma 3: Lemma 1 implies that $\theta \le \exp(-c\gamma^{-2r/5})$, where $c > 0$ is a constant depending only on r. Applying Lemma 8 and taking union bound over $k = \exp(c_1\gamma^{-2r/5})$ statistical queries, the proof follows.

4 Lower bounds on convex optimization

In this section, we describe the implications of our main lower bound learning of linear models with a convex loss. Consider the task of optimizing a convex function $\ell(w; P) := \mathbb{E}_{(x,y)\sim P} \ell(w; (x, y))$, where ℓ is a convex and Lipschitz *linear model*, namely, $\ell(w; (x, y)) = \varphi(\langle w, x\rangle y)$ for a function $\varphi\colon [-1, 1] \to \mathbb{R}$ which is convex

and Lipschitz and w is optimized over B_d, the unit ball in $\mathbb{R}^d$. Additionally, we will let $X = \mathbb{S}^{d-1}$ be the unit sphere in $\mathbb{R}^d$.We present two reductions: first, the standard reduction to *hinge loss* which is used in the *soft-margin support vector machine* (SVM) algorithm, and secondly, a reduction to a different function which is smooth and strongly convex, to show that a lower bound holds even given these assumptions.

In the first reduction, we set $\ell(w;(x,y)) = \max(0, \gamma - \langle w, x\rangle y)$. This loss has the nice property that if $w \in \mathbb{S}^{d-1}$ classifies (x,y) correctly with margin γ, namely, if $\langle w, x\rangle y \geq \gamma$, then $\ell(w;(w,y)) = 0$. At the same time, if w misclassifies (x,y), namely, $\langle w, x\rangle y \leq 0$, then $\ell(w;(x,y)) \geq \gamma$ (essentially, ℓ can be viewed as a scaled *surrogate loss function*). Thus, if the distribution P is linearly separable with margin γ, then $\min_{w \in B_d} \ell(w;P) = 0$, and any w satisfying $\ell(w;P) = \gamma/3$ is an approximate linear separator: $\Pr_{(x,y)\sim P}[\mathrm{sign}(\langle w, x\rangle) \neq y] \leq 1/3$. Hence, we can reduce solving linear models to classification, obtaining the following result:

Theorem 6. *For any $\alpha > 0$ there exists a loss function $\ell(w,(x,y)) = \varphi(y\langle w, x\rangle)$ where φ is convex and 1-Lipschitz, such that any non-interactive ϵ-LDP algorithm $\mathcal{A}$ that outputs $\widehat{w}$ satisfying $\mathbb{E}_{\mathcal{A}}[\ell(\widehat{w};P)] \leq \inf_{w\in B_d} \ell(w;P) + \alpha$, requires*

$$n \geq \min\left(\exp\left(\alpha^{-\Omega(1)}\right), \exp\left(d^{\Omega(1)}\right)\right)/e^{2\epsilon}.$$

Proof. Fix such algorithm $\mathcal{A}$. As discussed above, by simulating $\mathcal{A}$ we can approximately solve any classification problem with margin of at least 3α. Denote $\gamma = \max(3\alpha, d^{-1/(2+1/5)})$. Applying the lower bound on classification (Theorem 1), we derive that $n \geq \exp(\gamma^{-\Omega(1)})$, and the result follows. □

Before proceeding, we define smoothness and strong convexity:

Definition 4. *A differentiable function $f\colon [-1,1] \to \mathbb{R}$ with derivative $f'(x)$ is* σ-smooth *and μ-strongly convex for $0 \leq \mu \leq \sigma$, if for any $x, x' \in [-1,1]$,*

$$f'(x)(y-x) + \frac{\mu}{2}(y-x)^2 \leq f(y) - f(x) \leq f'(x)(y-x) + \frac{\sigma}{2}(y-x)^2.$$

Next, we define the following convex loss function $\varphi_\gamma\colon [-1,1] \to \mathbb{R}$:

$$\varphi_\gamma(t) = \frac{(1-t)^2}{8} + \begin{cases} 1 - 2t/\gamma & -1 \leq t \leq 0 \\ (t-\gamma)^2/\gamma^2 & 0 \leq t \leq \gamma \\ 0 & \gamma \leq t \leq 1. \end{cases}$$

Note that $\varphi_\gamma(t) - (1-t)^2/8$ is non-negative, monotonic non-decreasing, $2/\gamma$-Lipschitz, $2/\gamma^2$-smooth and convex. This implies that $\varphi_\gamma(t)$ is non-negative, monotonic non-decreasing, $2/\gamma + 1/2 \leq 3/\gamma$-Lipschitz, $2/\gamma^2 + 1/4 \leq 3/\gamma^2$-smooth and 1/4-strongly convex. Additionally, the following holds:

Claim 1. *For any $t \geq \gamma$, it holds that $\varphi_\gamma(t) \leq \varphi_\gamma(\gamma) \leq 1/8$, while for any $t \leq 0$, $\varphi_\gamma(t) \geq \varphi_\gamma(0) \geq 9/8$.*

Given a classification problem with distribution P over $X \times Y$ and margin γ, we reduce it to the convex optimization problem with the loss function $\ell_\gamma(w;(x,y)) = \varphi_\gamma(\langle w, x\rangle y)$. Claim 1 and Markov's inequality imply the following connection between the classification error and the convex loss:

Claim 2. *For any vector $w \in B_d$,*

$$\mathrm{err}_P(w) \leq \frac{\ell_\gamma(w;P)}{9/8} \leq \ell_\gamma(w;P),$$

where $\mathrm{err}_P(w)$ is the classification error of the function $x \mapsto \mathrm{sign}(\langle w, x\rangle)$. Additionally, $\inf_{w\in B_d} \ell_\gamma(w;P) \leq 1/8$ (in particular, any unit-norm vector that classifies correctly with γ-margin would achieve this loss).

Based on the above claim, we derive the following relationship between the expected error and expected loss:

Claim 3. *Let $\mathcal{A}$ be a (randomized) algorithm that for some distribution P over $\mathbb{S}^{d-1} \times \{-1,1\}$ outputs $\widehat{w} \in B_d$ that satisfies*

$$\mathop{\mathbb{E}}_{\mathcal{A}}[\ell_\gamma(\widehat{w}; P)] \leq \inf_{w \in B_d} \ell_\gamma(w; P) + 1/8.$$

Then, $\mathbb{E}_{\mathcal{A}}[\mathrm{err}_P(\widehat{w})] \leq 1/4$.

Proof. Applying both statements in Claim 2, we obtain that

$$\mathop{\mathbb{E}}_{\mathcal{A}}[\mathrm{err}_P(\widehat{w})] \leq \mathop{\mathbb{E}}_{\mathcal{A}}[\ell_\gamma(\widehat{w}; P)] \leq \inf_{w \in B_d} \ell_\gamma(w; P) + 1/8 \leq 1/4.$$

□

We are ready to state our lower bound for learning of a linear model with smooth and strongly convex loss:

Theorem 7. *For any parameters $0 \leq \mu < \sigma \leq \infty$, $L > 0$ and $\alpha > 0$, there exists a loss function $\ell(w, (x, y)) = \varphi(y\langle w, x\rangle)$ where φ is convex, L-Lipschitz, σ-smooth and μ-strongly convex, such that any non-interactive ϵ-LDP algorithm $\mathcal{A}$ that outputs $\widehat{w}$ satisfying $\mathbb{E}_{\mathcal{A}}[\ell(\widehat{w}; P)] \leq \inf_{w \in B_d} \ell(w; P) + \alpha$, requires*

$$n \geq \exp\left(c \min\left(\left(\frac{L}{\max(\mu, \alpha)}\right)^{2/5 \cdot (1-\xi)}, \left(\frac{\sigma}{\max(\mu, \alpha)}\right)^{1/5 \cdot (1-\xi)}, d^{1/6 \cdot (1-\xi)}\right)\right),$$

where ξ can be any number in $(0, 1)$ and $c > 0$ depends only on ξ.

Proof. In the proof we will allow redundancy in the parameters up to universal constants (e.g., requiring ℓ to be $\Omega(\mu)$ rather than μ strongly convex). Assume that $\max(\mu, \alpha) \leq \min(L, \sigma)$, otherwise the bound trivially follows. Denote $\theta = \max(\mu, \alpha)$ and let

$$\gamma := \max\left(\frac{\max(\mu, \alpha)}{L}, \left(\frac{\max(\mu, \alpha)}{\sigma}\right)^{1/2}, d^{-1/(2+2/5)}\right) = \max\left(\frac{\theta}{L}, \left(\frac{\theta}{\sigma}\right)^{1/2}, d^{-1/(2+2/5)}\right).$$

Consider the function $\theta\varphi_\gamma$: its Lipschitz constant is bounded by $3\theta/\gamma \leq 3L$, the smoothness parameter is bounded by $3\theta/\gamma^2 \leq 3\theta/(\sigma/\theta) = 3\sigma$ and the strong convexity parameter equals $\theta/4 \geq \mu/4$. Let $\mathcal{A}$ be an algorithm which finds an $\alpha/8$-optimal solution to the linear model defined by the function $\theta\varphi_\gamma$ (in expectation). Then, $\mathcal{A}$ finds an $\alpha/(8\theta) \leq 1/8$-optimal solution to optimization of linear models with loss defined by φ_γ. By Claim 3, $\mathcal{A}$ finds approximate linear separator to any classification problem with margin γ. Since γ is defined to satisfy $d \geq \gamma^{-2-2/5}$, by the lower bound on classification (Theorem 1), for any $r \in (0, 1)$, the sample complexity satisfies

$$n \geq \exp(c(\gamma^{-2r/5})),$$

where the constant $c > 0$ may depend only on r. Taking $r = 1 - \xi$, completes the proof. □

5 Implications for distributed learning with communication constraints

In this section we briefly define the model of bounded communication per sample, state the known equivalence results to the SQ model and spell out the immediate corollary of our lower bound. In the bounded communication model [BD98; SVW16] it is assumed that the total number of bits learned by the server about each data sample is bounded by ℓ for some $\ell \ll \log |Z|$. As in the case of LDP this is modeled by using an appropriate oracle for accessing the dataset. For simplicity we only introduce the non-interactive version of this model.

Definition 5. *We say that a (possibly randomized) algorithm $R\colon Z \to \{0,1\}^\ell$ extracts ℓ bits. For a dataset $S \in Z^n$, an COMM_S oracle takes as an input an index i and an algorithm R and outputs a random value w obtained by applying $R(z_i)$. A non-interactive algorithm is ℓ-bit communication bounded if it accesses S only via the ℓ-bit COMM_S oracle each sample is accessed once and all of its queries are determined before observing any of the oracle's responses.*

As first shown by Ben-David and Dichterman [BD98], it is easy to simulate a single query to COMM applied to a random sample from distribution P using a single query to $\mathrm{STAT}_P(\tau)$. The simulation has been strengthened in [Fel+12] and generalized to the COMM oracle that can access each sample more than once in [SVW16].

Theorem 8 ([SVW16])**.** *Let $\mathcal{A}$ be a non-interactive ℓ-bit communication bounded algorithm that makes queries to COMM_S for $S \in Z^n$ drawn i.i.d. from some distribution P. Then for every $\delta > 0$, there is an SQ algorithm $\mathcal{A}_{SQ}$ that makes $2n\ell$ non-adaptive queries to $STAT_P\left(\delta/(2^{\ell+1}n)\right)$ and produces the same output as $\mathcal{A}$ with probability at least $1-\delta$.*

A direct corollary of Theorems 5 and 8 is the following lower bound:

Corollary 1. *Fix $\gamma \in (0,1/2)$, $r \in (0,1)$ and $d \geq \gamma^{-2-2r/5}$. Let $\mathcal{A}$ be a non-interactive ℓ-bit communication bounded algorithm with n users. Assume that for any classification problem (D, f^*) over $\mathbb{R}^d$ with margin $\gamma(f^*, D) \geq \gamma$, the algorithm outputs a hypothesis $\widehat{f}$ with expected loss $\mathbb{E}_{\mathcal{A}}[\mathrm{err}_{f^*,D}(\widehat{f})] \leq 1/2 - \gamma^{1-r}$. Then, either $\ell \geq c\gamma^{-2r/5}$ or $n \geq \exp(c\gamma^{-2r/5})$, where $c > 0$ is a constant depending only on r.*

The lower bound for learning linear models with convex loss can be extended analogously.

A Total variation distance and data processing inequality

Let P, Q be two probability distributions defined on the same probability space $(\Omega, \mathcal{F})$, where Ω is the sample space and $\mathcal{F}$ is the set of events. The *total variation distance* between P and Q is defined as

$$d_{\mathrm{TV}}(P,Q) := \max_{F \in \mathcal{F}} |P(F) - Q(F)|.$$

Given two spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, a *random function* f from Ω_1 to Ω_2 is a function that given $a \in \Omega_1$, outputs a probability distribution f_a over Ω_2 (ignoring measurability issues). Given a probability distribution P over Ω_1, we define $f(P)$ as the probability distribution obtained by drawing $a \sim P$, drawing $b \sim f_a$ and then outputting b. The *data processing inequality* states the following:

Lemma 11 (Data processing inequality)**.** *Let P and Q be two distributions over Ω_1 and let f be a measurable random function from Ω_1 to Ω_2. Then,*

$$d_{\mathrm{TV}}(f(P), f(Q)) \leq d_{\mathrm{TV}}(P,Q).$$

This lemma is intuitive: if P and Q are similar, applying the same transformation f cannot make them different.

References

[ABR64] M. A. Aizerman, E. A. Braverman, and L. Rozonoer. "Theoretical foundations of the potential function method in pattern recognition learning." In: *Automation and Remote Control,* Automation and Remote Control, 25. 1964, pp. 821–837.

[Ach+19] J. Acharya, C. De Sa, D. Foster, and K. Sridharan. "Distributed Learning with Sublinear Communication". In: *International Conference on Machine Learning*. 2019, pp. 40–50.

[ACT18] J. Acharya, C. L. Canonne, and H. Tyagi. "Inference under Information Constraints I: Lower Bounds from Chi-Square Contraction". In: *arXiv preprint arXiv:1812.11476* (2018).

[ACT19] J. Acharya, C. L. Canonne, and H. Tyagi. "Inference under information constraints II: Communication constraints and shared randomness". In: *arXiv preprint arXiv:1905.08302* (2019).

[AK65] N. I. Akhiezer and N Kemmer. *The classical moment problem: and some related questions in analysis*. Vol. 5. Oliver & Boyd Edinburgh, 1965.

[App17] Apple's Differential Privacy Team. "Learning with Privacy at Scale". In: *Apple Machine Learning Journal* 1.9 (Dec. 2017).

[ASZ19] J. Acharya, Z. Sun, and H. Zhang. "Hadamard Response: Estimating Distributions Privately, Efficiently, and with Little Communication". In: *The 22nd International Conference on Artificial Intelligence and Statistics*. 2019, pp. 1120–1129.

[BD98] S. Ben-David and E. Dichterman. "Learning with Restricted Focus of Attention". In: *J. Comput. Syst. Sci.* 56.3 (1998), pp. 277–298.

[BES02] S. Ben-David, N. Eiron, and H. U. Simon. "Limitations of Learning Via Embeddings in Euclidean Half Spaces". In: *Journal of Machine Learning Research* 3 (2002), pp. 441–461.

[BGGP12] I. Benjamini, O. Gurel-Gurevich, and R. Peled. "On k-wise independent distributions and boolean functions". In: *arXiv preprint arXiv:1201.3261* (2012).

[BGV92] B. E. Boser, I. Guyon, and V. Vapnik. "A Training Algorithm for Optimal Margin Classifiers". In: *COLT*. ACM, 1992, pp. 144–152.

[Blu+97] A. Blum, A. Frieze, R. Kannan, and S. Vempala. "A polynomial time algorithm for learning noisy linear threshold functions". In: *Algorithmica* 22.1/2 (1997), pp. 35–52.

[BNS18] M. Bun, J. Nelson, and U. Stemmer. "Heavy Hitters and the Structure of Local Privacy". In: *Symposium on Principles of Database Systems*. 2018, pp. 435–447.

[BS15] R. Bassily and A. Smith. "Local, Private, Efficient Protocols for Succinct Histograms". In: *STOC* (2015).

[BS18] E. Balkanski and Y. Singer. "Parallelization does not Accelerate Convex Optimization: Adaptivity Lower Bounds for Non-smooth Convex Minimization". In: *CoRR* abs/1808.03880 (2018). arXiv: 1808.03880.

[DF18] A. Daniely and V. Feldman. "Locally private learning without interaction requires separation". In: *arXiv preprint arXiv:1809.09165* (2018). Extended abstract in NeurIPS 2019.

[DF19] A. Daniely and V. Feldman. "Open Problem: Is Margin Sufficient for Non-Interactive Private Distributed Learning?" In: *COLT*. 2019, pp. 3180–3184.

[DG18] J. Diakonikolas and C. Guzmán. "Lower Bounds for Parallel and Randomized Convex Optimization". In: *CoRR* abs/1811.01903 (2018). arXiv: 1811.01903.

[DJW13] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. "Local Privacy and Statistical Minimax Rates". In: *FOCS*. 2013, pp. 429–438.

[DKY17] B. Ding, J. Kulkarni, and S. Yekhanin. "Collecting Telemetry Data Privately". In: *31st Conference on Neural Information Processing Systems (NIPS)*. 2017, pp. 3574–3583.

[DR19] J. Duchi and R. Rogers. "Lower Bounds for Locally Private Estimation via Communication Complexity". In: *arXiv preprint arXiv:1902.00582* (2019).

[DRY18] J. C. Duchi, F. Ruan, and C. Yun. "Minimax Bounds on Stochastic Batched Convex Optimization". In: *COLT*. 2018, pp. 3065–3162.

[DWJ13] J. C. Duchi, M. J. Wainwright, and M. I. Jordan. "Local Privacy and Minimax Bounds: Sharp Rates for Probability Estimation". In: *NIPS*. 2013, pp. 1529–1537.

[Dwo+06] C. Dwork, F. McSherry, K. Nissim, and A. Smith. "Calibrating noise to sensitivity in private data analysis". In: *TCC*. 2006, pp. 265–284.

[EGS03] A. V. Evfimievski, J. Gehrke, and R. Srikant. "Limiting privacy breaches in privacy preserving data mining". In: *PODS*. 2003, pp. 211–222.

[EPK14] Ú. Erlingsson, V. Pihur, and A. Korolova. "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response". In: *ACM SIGSAC Conference on Computer and Communications Security*. 2014, pp. 1054–1067.

[Erl+18] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta. "Amplification by Shuffling: From Local to Central Differential Privacy via Anonymity". In: *CoRR* abs/1811.12469 (2018). Extended abstract in SODA 2019. arXiv: 1811.12469.

[Fel+12] V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. "Statistical Algorithms and a Lower Bound for Detecting Planted Cliques". In: *arXiv, CoRR* abs/1201.1214 (2012). Extended abstract in STOC 2013.

[FGV15] V. Feldman, C. Guzman, and S. Vempala. "Statistical Query Algorithms for Mean Vector Estimation and Stochastic Convex Optimization". In: *CoRR* abs/1512.09170 (2015). Extended abstract in SODA 2017.

[FSS01] J. Forster, N. Schmitt, and H. U. Simon. "Estimating the Optimal Margins of Embeddings in Euclidean Half Spaces". In: *Proceedings of COLT 2001 and EuroCOLT 2001*. 2001, pp. 402–415.

[HKR12] J. Hsu, S. Khanna, and A. Roth. "Distributed Private Heavy Hitters". In: *ICALP*. 2012, pp. 461–472.

[JMR19] M. Joseph, J. Mao, and A. Roth. "Exponential Separations in Local Differential Privacy Through Communication Complexity". In: *CoRR* abs/1907.00813 (2019). arXiv: 1907.00813.

[Jos+19] M. Joseph, J. Mao, S. Neel, and A. Roth. "The Role of Interactivity in Local Differential Privacy". In: *CoRR* abs/1904.03564 (2019). arXiv: 1904.03564.

[Kas+11] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. "What Can We Learn Privately?" In: *SIAM J. Comput.* 40.3 (June 2011), pp. 793–826.

[Kea98] M. Kearns. "Efficient noise-tolerant Learning from statistical queries". In: *Journal of the ACM* 45.6 (1998), pp. 983–1006.

[KN77] M. G. Krein and A. A. Nudel'man. *The Markov moment problem and extremal problems: ideas and problems of PL Cebysev and AA Markov and their further development*. American Mathematical Society, 1977.

[KS11] M. Kallweit and H. Simon. "A Close Look to Margin Complexity and Related Parameters". In: *COLT*. 2011, pp. 437–456.

[LS09] N. Linial and A. Shraibman. "Learning Complexity vs Communication Complexity". In: *Comb. Probab. Comput.* 18.1-2 (Mar. 2009), pp. 227–245. ISSN: 0963-5483.

[Luo05] Z.-Q. Luo. "Universal decentralized estimation in a bandwidth constrained sensor network". In: *IEEE Transactions on information theory* 51.6 (2005), pp. 2210–2219.

[Nov62] A. Novikoff. "On convergence proofs on perceptrons". In: *Proceedings of the Symposium on Mathematical Theory of Automata*. Vol. XII. 1962, pp. 615–622.

[RG06] A. Ribeiro and G. B. Giannakis. "Bandwidth-constrained distributed estimation for wireless sensor networks-part I: Gaussian case". In: *IEEE transactions on signal processing* 54.3 (2006), pp. 1131–1143.

[RWV06] R. Rajagopal, M. J. Wainwright, and P. Varaiya. "Universal quantile estimation with feedback in the communication-constrained setting". In: *Information Theory, 2006 IEEE International Symposium on*. IEEE. 2006, pp. 836–840.

[SD15] J. Steinhardt and J. C. Duchi. "Minimax rates for memory-bounded sparse linear regression". In: *COLT*. 2015, pp. 1564–1587.

[She08] A. A. Sherstov. "Halfspace Matrices". In: *Computational Complexity* 17.2 (2008), pp. 149–178.

[STU17] A. D. Smith, A. Thakurta, and J. Upadhyay. "Is Interaction Necessary for Distributed Private Learning?" In: *2017 IEEE Symposium on Security and Privacy, SP 2017*. 2017, pp. 58–77.

[Sur+16] A. T. Suresh, F. X. Yu, H. B. McMahan, and S. Kumar. "Distributed mean estimation with limited communication". In: *arXiv preprint arXiv:1611.00429* (2016).

[SVW16] J. Steinhardt, G. Valiant, and S. Wager. "Memory, Communication, and Statistical Queries". In: *COLT*. 2016, pp. 1490–1516.

[SWv] A. Smith, and D. Wang. Personal communication. Nov, 2019.

[Ull18] J. Ullman. "Tight Lower Bounds for Locally Differentially Private Selection". In: *CoRR* abs/1802.02638 (2018). arXiv: 1802.02638.

[Wan+19] D. Wang, H. Zhang, M. Gaboardi, and J. Xu. "Estimating Smooth GLM in Non-interactive Local Differential Privacy Model with Public Unlabeled Data". In: *arXiv preprint arXiv:1910.00482* (2019).

[War65] S. L. Warner. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias". In: *J. of the American Statistical Association* 60.309 (1965), pp. 63–69.

[WGX18] D. Wang, M. Gaboardi, and J. Xu. "Empirical Risk Minimization in Non-interactive Local Differential Privacy Revisited". In: *Advances in Neural Information Processing Systems 31*. 2018, pp. 965–974.

[Woo+18] B. E. Woodworth, J. Wang, A. D. Smith, B. McMahan, and N. Srebro. "Graph Oracle Models, Lower Bounds, and Gaps for Parallel Stochastic Optimization". In: *NeurIPS*. 2018, pp. 8505–8515.

[WSX19] D. Wang, A. Smith, and J. Xu. "Noninteractive Locally Private Learning of Linear Models via Polynomial Approximations". In: *Proceedings of the 30th International Conference on Algorithmic Learning Theory*. Vol. 98. Proceedings of Machine Learning Research. 2019, pp. 898–903.

[WX19] D. Wang and J. Xu. "Principal component analysis in the local differential privacy model". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*. 2019, pp. 4795–4801.

[Zha+13] Y. Zhang, J. C. Duchi, M. I. Jordan, and M. J. Wainwright. "Information-theoretic lower bounds for distributed statistical estimation with communication constraints". In: *NIPS*. 2013, pp. 2328–2336.

[ZMW17] K. Zheng, W. Mou, and L. Wang. "Collect at once, use effectively: making non-interactive locally private learning possible". In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org. 2017, pp. 4130–4139.