# Estimating Euclidean distance to linearity

Andrej Bogdanov Lorenzo Taschin

**Abstract**

Given oracle access to a real-valued function on the $n$-dimensional Boolean cube, how many queries does it take to estimate the squared Euclidean distance to its closest linear function within $\epsilon$? Our main result is that $O(\log^3(1/\epsilon) \cdot 1/\epsilon^2)$ queries suffice. Not only is the query complexity independent of $n$ but it is optimal up to the polylogarithmic factor.

Our estimator evaluates $f$ on pairs correlated by noise rates chosen to cancel out the low-degree contributions to $f$ while leaving the linear part intact. The query complexity is optimized when the noise rates are multiples of Chebyshev nodes.

In contrast, we show that the dependence on $n$ is unavoidable in two closely related settings. For estimation from random samples, $\Theta(\sqrt{n}/\epsilon + 1/\epsilon^2)$ samples are necessary and sufficient. For agnostically learning a linear approximation with $\epsilon$ mean-square regret under the uniform distribution, $\Omega(n/\sqrt{\epsilon})$ nonadaptively chosen queries are necessary, while $O(n/\epsilon)$ random samples are known to be sufficient (Linial, Mansour, and Nisan).

Our upper bounds apply to functions with bounded 4-norm. Our lower bounds apply even to $\pm 1$-valued functions.

## 1 Introduction

Finding linear approximations is perhaps the most important problem in statistics and data science. *Linear regression* [11, Section 9.2] seeks to learn the "best" linear predictor $\ell$ for $f$ from labeled data $(x, f(x))$. Owing to its wide applicability and computational advantages, approximation error is often measured as the minimum squared loss $\mathbf{E}[(f(x) - \ell(x))^2]$ under some distribution on examples.

Our main question of interest is whether *estimating* the minimum squared loss can be performed more efficiently than learning the predictor. A highly efficient estimator can be potentially used to decide whether learning a linear model is sensible at all.

We aim to highlight the conceptual distinctions between estimation and learning. We demonstrate that a complexity gap between the two problems is already exhibited by arguably the simplest distribution on examples: The uniform distribution over the Boolean cube $\{\pm 1\}^n$.

Under the uniform distribution, the minimum squared loss is the distance between the function of interest viewed as a vector in $2^n$-dimensional Euclidean space and the $n$-dimensional subspace spanned by the linear functions. In this setting estimating the squared distance is essentially equivalent to estimating the squared length of the projection.

The power of linear regression in machine learning applications is derived to a great extent from the distribution-free nature of the training algorithms. In contrast our results concern linear regression for a known distribution. We believe that distribution-specific regression is interesting for several reasons.

First, restrictive assumptions on the data distribution may improve performance. For example, distributions in which some features linearly approximate other features can be problematic for popular algorithms like gradient descent. The uniform distribution, as a model of pairwise independence, avoids such obstacles.

Second, distribution-specific algorithms sometimes serve as a stepping stone to distribution-free ones. In the context of property testing this strategy was successful in obtaining distribution-free algorithms for combinatorial properties like monotonicity [6] as well as algebraic ones like linearity and low-degree representability [6, 5].

Third, negative results about learning and estimation are stronger in the distribution-specific model. Distribution-free learning must succeed for all concepts in the given class *and* all distributions on examples. To understand the limitations it is natural to decouple the complexities of the concept class and of the distribution. Do learning and estimation remain hard even when the distribution is simple?

In this work we focus on information-theoretic measures of complexity. Our principal measure of performance is the *query complexity* $m$. The time complexity of all our algorithms is linear in $mn$, namely in the size of the labeled dataset. We study both algorithms that choose their queries and ones that are provided with uniformly random examples.

## Our results

Our results are summarized in Table 1. Our main contribution is Theorem 3.1: The squared linear projection of $f$

$$\mathbf{W}^1[f] = \mathbf{E}[f^2] - \min_{\text{linear } \ell} \mathbf{E}[(f(x) - \ell(x))^2] \tag{1}$$

can be estimated within $\epsilon$ with $\tilde{O}(1/\epsilon^2)$ chosen queries.[1] In contrast, learning the requisite approximation and obtaining the same estimate from random samples both entail queries that grow at least as fast as some root of $n$.

| | chosen queries | random samples |
|---|---|---|
| estimation | $O(\log^3(1/\epsilon) \cdot 1/\epsilon^2)$ (Theorem 3.1)<br>$\Omega(1/\epsilon^2)$ (Theorem 3.8) | $O(\sqrt{n}/\epsilon + 1/\epsilon^2)$ (Theorem 5.1)<br>$\Omega(\sqrt{n}/\epsilon + 1/\epsilon^2)$ (Theorem 5.2 + 3.8) |
| learning | $O(n/\epsilon)$ [8]<br>$\Omega(n/\sqrt{\epsilon})$ non-adaptive (Theorem 4.1) | $O(n/\epsilon)$ [8]<br>$\Omega(n/\sqrt{\epsilon})$ (Theorem 4.1) |

Table 1: Query complexity of $\epsilon$-estimating $\mathbf{W}^1[f]$ and $\epsilon$-learning the linear part of $f$. All positive results assume $\|f\|_4 \leq 1$. All negative results hold even for $\pm 1$-valued $f$.

We found this result surprising from a Fourier-analytic perspective. Our quantity of interest is the sum of squares of all $n$ first-level Fourier coefficients of $f$. Up to the polylogarithmic factor, estimating the sum has the same query complexity as estimating any of its individual terms, or estimating the squared mean of $f$.

In the decision version of estimation also known as *tolerant testing* [10], the objective is to distinguish functions that are $\theta$-close to linear from those that are $\theta + \epsilon$-far. In particular, the algorithm of Theorem 3.1 is a tolerant tester for any value of $\theta$. In the intolerant limit $\theta = 0$, algorithms with $O(1/\epsilon)$ query complexity are known [1, 3]. For functions over $\mathbb{R}^n$ the query complexity is also $O(1/\epsilon)$ under Gaussian measure [7], and $O(\log(1/\epsilon) \cdot 1/\epsilon)$ under arbitrary measure [5]. With the exception of [7], these testers appear not to be robust against small perturbations as their soundness is analyzed with respect to the Hamming distance. The tester of Khot and Moshkovitz [7] can be viewed as an estimator, albeit one with large constant bias.

Our positive results apply to functions with bounded 4-norm. Boundedness of the 2-norm is clearly insufficient: The point function $f(x) = 2^n \mathbb{1}(x = a)$ for a random point $a$ has unit 2-norm and unit squared distance to linearity, yet is indistinguishable from the all-zero function with $o(2^n)$ queries. We did not investigate whether $p$-norm boundedness for $2 < p < 4$ is sufficient.

In contrast, in Theorem 4.1 we show that proper learning of a linear hypothesis for $f$ that is within $\epsilon$ squared distance of optimal has non-adaptive query complexity linear in $n$, specifically $\Omega(n/\sqrt{\epsilon})$. Our technique also yields a $\Omega(\log n/\sqrt{\epsilon})$ general (adaptive) lower bound. On the positive side, it is known that $O(n/\epsilon)$ random samples are sufficient for learning via empirical risk mininization [8]. The $\sqrt{\epsilon}$ gap between the lower and upper bounds is reminiscent of analogous gaps in distribution-free agnostic learning of deterministic concept classes [2].

In Theorems 5.1 and 5.2 we show that the sample complexity of estimating $\mathbf{W}^1[f]$ is $\sqrt{n}/\epsilon + 1/\epsilon^2$ up to constant factor. As a consequence of Theorems 5.1 and 4.1, estimation requires fewer random samples than learning in the regime $\epsilon = \omega(n^{-2/3})$.

## Techniques

**Estimation from chosen queries** Algorithm 1 estimates the linear projection of $f$ by a suitably scaled product $f(x)f(y)$ evaluated on a pair $(x, y)$ of correlated inputs. The query complexity can be

[1] The query complexities of the squared projection and the squared distance to linearity are the same up to constant factor since $\mathbf{E}[f^2]$ can be estimated from $O(1/\epsilon^2)$ samples assuming $\|f\|_4$ is bounded. For similar reasons so are the squared projection and distance to the space of affine functions.

decoupled from the input length $n$ already when the pairs $(x_i, y_i)$ are independent, marginally uniform, and of sufficiently low correlation $\rho = \mathbf{E}\, x_i y_i$. For simplicity assume $f$ is unbiased and $\pm 1$-valued. By orthogonal decomposition (see (2)),

$$\mathbf{E}\, f(x)f(y) = \rho\|f^{=1}\|^2 + \rho^2\|f^{=2}\|^2 + \cdots + \rho^n\|f^{=n}\|^2,$$

where $f^{=d}$ is the degree-$d$ part of $f$ (see precise definition in Section 2). By Parseval's identity, $\rho^{-1}f(x)f(y)$ estimates $\mathbf{W}^1[f] = \|f^{=1}\|^2$ with bias at most $\sum_{k>1}|\rho|^{k-1}\|f^{=k}\|^2 \le |\rho|$. By choosing the bias-variance tradeoff $\rho \approx \epsilon/2$, empirically averaging this estimator yields an $\epsilon$-approximation of $\mathbf{W}^1[f]$ with $O(1/\epsilon^4)$ queries.

To improve the query complexity we work with a more general class of distributions on pairs $(x, y)$. Barring additional information on $f$, it is sensible that the distribution should place equal probability on all pairs $(x, y)$ whose difference has fixed Hamming weight. A special class of such distibutions is mixtures of $\rho$-correlated pairs over some distribution $d(\rho)$ on noise rates.

The uniform mixture of $\rho$ and $-\rho$-correlated pairs already improves the query complexity to $O(1/\epsilon^3)$. The reason is that the related Fourier decomposition

$$\mathbf{E}\ (\text{sign of correlation})f(x)f(y) = \rho\|f^{=1}\|^2 + \rho^3\|f^{=3}\|^2 + \rho^5\|f^{=5}\|^2 + \cdots$$

cancels out all even levels, thus reducing the bias of the estimator from $\rho$ to $\rho^2$.

To achieve our query complexity of $O(\log^3(1/\epsilon)1/\epsilon^2)$ we construct a mixture of noise rates $\alpha_1\rho, \ldots, \alpha_\ell\rho$ that annihilates levels 2 up to $\ell - 1$ while minimizing the contribution of levels $\ell$ and higher. A close to optimal error is achieved using Chebyshev interpolation. We give evidence that a $\Omega(\log^2(1/\epsilon)1/\epsilon^2)$ lower bound is inherent in this approach (Claim 3.6), but we do not know if it can be bypassed in general.

In Remark 3.7 we explain why the low complexity of Algorithm 1 owes not only to the choice of interpolation scheme, but to some serendipity in the constraints (6) on the noise rates.

The proof of Theorem 3.8 is based on a reduction from the problem of estimating the bias of a coin.

**Learning lower bound** Theorem 4.1 is proved in two steps. Proposition 4.2 addresses the regime of constant error $\epsilon$. It is proved by exhibiting a family of $2^{\Omega(n)}$ functions whose linear projections are $\Omega(1)$-far apart. The existence of this family is established using the probabilistic method in Lemma 4.3. An algorithm of sublinear query complexity is unlikely to disambiguate among members of the family and cannot be accurate.

In Proposition 4.5 we give a self-reduction for learning linear approximations from query complexity $q$ and constant error to query complexity $O(q/\sqrt{\epsilon})$ and error $\epsilon$. The reduction applies to any query model.

**Estimation from samples** The random variable $f(x)f(y)\langle x, y\rangle$ with random $x, y$ is an unbiased estimator of $\mathbf{W}^1[f]$. To improve its variance at optimal query cost we average it over all $\binom{m}{2}$ pairs of examples (Algorithm 2 and Theorem 5.1). For constant $\epsilon$, the lower bound in Theorem 5.2 is exhibited by a sign-rounded linear function $\ell$ with independent Gaussian coefficients. Such a function is weakly pseudorandom against $o(\sqrt{n})$ examples yet has constant correlation with $\ell$. We achieve optimal dependence on $\epsilon$ by adding Gaussian noise to $\ell$.

## Extensions and future work

We believe that a result analogous to Theorem 3.1 can be obtained for functions over $\mathbb{R}^n$ under standard Gaussian measure using related techniques. It would be interesting to investigate the general setting of product distributions under Efron-Stein decomposition. The biased Boolean cube could be a good test case.

Theorem 3.1 can be extended to approximate the degree-$d$ part of $f$ with $\log^{O(d)}(1/\epsilon)\cdot 1/\epsilon^2$ queries. Both extensions might be sensible in the context of regression, where a mix of categorical (Boolean) and numerical (Gaussian) data and higher-degree models (capturing e.g. decision trees) are more realistic.

An interesting open question is whether the polylogarithmic factor can be eliminated. We speculate that this may be possible under the stronger assumption that $f$ is bounded in infinity-norm.

### Organization

Section 2 has a brief background on Fourier analysis of Boolean functions. In Sections 3, 4, and 5 we present our results on estimation from chosen queries, learning, and estimation from samples, respectively.

## 2 Fourier analysis over the Boolean cube

We use standard notation from Boolean function analysis [9]. All functions are real-valued over the Boolean cube $\{\pm 1\}^n$ under uniform measure. Such functions live in Hilbert space under inner product $\mathbf{E} fg$. The $p$-norm of a function is $\|f\|_p = \mathbf{E}[|f|^p]^{1/p}$, with $p = 2$ usually omitted. The monomials $\{\prod_{i\in S} x_i : S \subseteq [n]\}$ form an orthonormal basis. $f$ admits orthogonal decompositions

$$f = f^{=0} + f^{=1} + \cdots + f^{=n} = \mathbf{E} f + f^{=1} + f^{\geq 2}$$

with $f^{=j}$ (resp., $f^{\geq j}$) being the projection onto the subspace spanned by monomials of degree exactly (resp., at least) $j$. Our main object of interest is the linear part $f^{=1}$, especially its weight

$$\mathbf{W}^1[f] = \|f^{=1}\|^2 = \mathbf{E} f f^{=1}.$$

The orthogonal decomposition can be further refined into the complete Fourier decomposition

$$f(x) = \sum\nolimits_{S\subseteq[n]} \hat{f}(S) \prod\nolimits_{i\in S} x_i, \qquad \text{where} \qquad \hat{f}(S) = \mathbf{E} f(x) \prod\nolimits_{i\in S} x_i.$$

By orthogonality, $\mathbf{E} fg = \sum \mathbf{E} f^{=i} g^{=i} = \sum \hat{f}(S)\hat{g}(S)$ (Plancherel's formula). In particular, when $f = g$

$$\|f\|^2 = \sum \|f^{=j}\|^2 = \sum \hat{f}(S)^2. \qquad \text{(Parseval's identity)}$$

By orthogonality, the distance to linearity $\mathbf{E}[(f(x)-\ell(x))^2]$ is minimized when $\ell = f^{=1}$. The minimizer equals

$$\min_{\text{linear } \ell} \mathbf{E}[(f(x) - \ell(x))^2] = \|f\|^2 - \|f^{=1}\|^2 = \mathbf{E}[f^2] - \mathbf{W}^1[f]$$

as in (1).

A $\rho$-biased string in $\{\pm 1\}^n$ is one in which the coordinates are independent and $\rho$-biased. We use $x \cdot y$ for the pointwise product of strings $x$ and $y$ in $\{\pm 1\}^n$. Multiplying by a $\rho$-biased string and averaging the outcome has the effect of dampening the higher-degree terms:

$$\mathbf{E}_{\rho\text{-biased } e}[f(x \cdot e)] = f^{=0}(x) + \rho f^{=1}(x) + \rho^2 f^{=2}(x) + \cdots \tag{2}$$

## 3 Estimation from chosen queries

**Theorem 3.1.** *There is an algorithm that makes $O(\log(\|f\|_4^2/\epsilon)^3 \cdot \|f\|_4^4/\epsilon^2)$ queries to $f$ and outputs a value within $\epsilon$ of $\mathbf{W}^1[f]$ with probability at least $2/3$.*

The algorithm takes the empirical average of $m$ instantiations of Algorithm 1 with parameters $m = 96\ell^3/\epsilon'^2$, $\ell = \log 1/\epsilon' + 1$, $\rho = \epsilon'^{(\ell-1)^{-1}}/6$, and $\epsilon' = \epsilon/\|f\|_4^2$. (We may assume $\epsilon'$ is an inverse power of two as this doesn't change the asymptotics.)

**Algorithm 1** Estimator $\tilde{W}^1[f]$ from chosen queries

**Parameters:** $0 < \rho < 1$ and $\ell \in \mathbb{N}$
**Setup:** Set $\alpha_1, \ldots, \alpha_\ell \in [-1, 1]$ as in (5). Calculate $d_1, \ldots, d_\ell$ as in (6).
Sample $i$ from $\{1, \ldots, \ell\}$ with probability $|d_i|/\|d\|_1$.
Sample random $x$ and $\alpha_i\rho$-biased $e_i$.
Output $\tilde{W}^1[f] = \rho^{-1}\|d\|_1 \cdot \operatorname{sign} d_i \cdot f(x)f(x \cdot e_i)$.

The conditional bias of this estimator given $x$ is $f(x)g(x)$, where

$$g = \rho^{-1} \sum d_i N_{\alpha_i\rho} f(x).$$

Here $N_\rho$ is the noise operator $N_\rho f(x) = \mathbf{E}[f(x \cdot e)]$ for a $\rho$-biased $e$. The constants $d_i, \alpha_i$ are chosen so that the first Fourier level of $f$ survives but all other levels among the first $\ell - 1$ are annihilated:

$$g^{=j} = \begin{cases} 0, & \text{if } j = 0 \text{ or } 2 \le j < \ell \\ f^{=j}, & \text{if } j = 1 \end{cases} \tag{3}$$

The theorem is proved by balancing the bias and the variance of the estimator $\tilde{W}^1$. The next two claims bound the bias and the variance respectively. Their proofs are in Section 3.1.

**Claim 3.2.** $|\mathbf{E}\,\tilde{W}^1[f] - \mathbf{W}^1[f]| \le \|d\|_1 \cdot \rho^{\ell-1} \|f^{\ge \ell}\|^2$.

**Claim 3.3.** $\tilde{W}^1[f]$ *has variance at most* $\rho^{-2} \cdot \|d\|_1^2 \cdot \|f\|_4^4$.

The quantity $\|d\|_1$ governs both. We analyze it next. Expanding $g$ in the Fourier basis and applying (2), the constraints (3) translate to

$$\rho^{-1} \sum_{i=1}^{\ell} d_i (\alpha_i \rho)^j = \begin{cases} 0, & \text{if } j = 0 \text{ or } 2 \le j < \ell \\ 1, & \text{if } j = 1. \end{cases}$$

We seek a short solution to the linear system

$$\sum_i d_i \alpha_i^j = \begin{cases} 0, & \text{if } j = 0 \text{ or } 2 \le j < \ell \\ 1, & \text{if } j = 1. \end{cases} \tag{4}$$

in unknowns $d_1, \ldots, d_\ell$. The existence of a short solution strongly depends on the choice of evaluation points $\alpha_1, \cdots, \alpha_\ell$. The Chebyshev nodes

$$\alpha_i = \cos\Big(\pi \frac{2i-1}{2\ell}\Big) \tag{5}$$

are close to best possible for this purpose, as will be argued in Corollary 3.5 and Claim 3.6.

**Claim 3.4.** $d_i = d(\alpha_i)$, *where*

$$d(t) = \frac{2}{\ell} \sum_{odd\ j\ =\ 1}^{\ell-1} (-1)^{(j-1)/2} \cdot j T_j(t), \tag{6}$$

*and* $T_j$ *is the* $j$*-th Chebyshev polynomial of the first kind.*

*Proof.* The first $\ell$ Chebyshev polynomials $T_0, \ldots, T_{\ell-1}$ are orthogonal with respect to inner product over the Chebyshev nodes:

$$\sum_i T_j(\alpha_i) T_{j'}(\alpha_i) = \begin{cases} 0, & \text{if } j \ne j', \\ \ell/2, & \text{if } j = j' \ne 0, \\ \ell, & \text{if } j = j' = 0. \end{cases}.$$

Using the facts $T_0(t) = 1, T_1(t) = t$, and that $T_j'(0)$ is zero for even $j$ and $(-1)^{(j-1)/2} j$ for odd $j$ we derive the unique Chebyshev expansion

$$d(t) = \sum \hat{d}(j) T_j(t). \tag{7}$$

By orthogonality, $\hat{d}(j)$ equals $(2/\ell) \sum d(\alpha_i) T_j(\alpha_i)$, when $j \ne 0$ and half that when $j = 0$. In the latter case, the first equation in (4) says that

$$\hat{d}(0) = \frac{1}{\ell} \sum_i d(\alpha_i) T_0(\alpha_i) = \frac{1}{\ell} \sum d_i = 0.$$

Otherwise, the polynomial $T_j(t) - T_j'(0)t$ has no linear term, so

$$\sum_i d(\alpha_i)(T_j(\alpha_i) - T_j'(0)\alpha_i)$$

expands as a linear combination of $\sum_i d_i \alpha_i^k$ for $k \le j, k \ne 1$. By (4) such linear combinations vanish. Therefore if $j \ne 1$,

$$\hat{d}(j) = \frac{2}{\ell} \sum_i d_i T_j'(0) \alpha_i = \frac{2}{\ell} T_j'(0) = \begin{cases} (2/\ell)(-1)^{(j-1)/2} \cdot j, & \text{if } j \text{ is odd} \\ 0, & \text{if } j \text{ is even.} \end{cases}$$

Plugging into (7) produces the desired formula (6). □

**Corollary 3.5.** $\|d\| \le \ell$.

*Proof.* By orthogonality we have the explicit formula

$$\|d\|^2 = \ell \hat{d}(0)^2 + \frac{\ell}{2} \hat{d}(1)^2 + \cdots + \frac{\ell}{2} \hat{d}(\ell-1)^2 = \frac{2}{\ell} \sum_{\text{odd } j=1}^{\ell-1} j^2,$$

which is at most twice the sum of odd integers between 1 and $\ell$. Those can be matched into $\ell/2$ pairs that add up to $\ell$ giving the $\ell^2$ upper bound. □

Therefore $\|d\|_1 \le \ell^{3/2}$. This bound is within a $\sqrt{\ell}$ factor of optimal for any choice of roots:

**Claim 3.6.** *For any choice of $-1 \le \alpha_1, \ldots, \alpha_\ell \le 1$, any $d$ that solves (4) has length $\|d\|_1 \ge \ell - 1$.*

Matching this bound (up to constant factor) would result in a $\log \|f\|_4^2/\epsilon$ factor improvement in query complexity. We do not know if $d$ as in (6) achieves it already.

*Proof.* Set $j$ equal to $\ell$ or $\ell - 1$, whichever is odd. Take a linear combination of equations (4) weighted by the coefficients $t$ of $T_j$. As $T_j$ is bounded on $[-1, 1]$,

$$\|d\|_1 \ge \left| \sum\nolimits_i d_i T_j(\alpha_i) \right| = \left| \sum\nolimits_{i,k} d_i \alpha_i^k t_k \right| = |t_1| = j.$$ □

*Remark* 3.7. By the same argument, had the right-hand side of (6) been replaced by $\mathbb{1}(j = \ell)$, there would have been no solution $d$ of 1-norm less than $2^{\ell-1}$. Thus the low complexity of Algorithm 1 owes not only to the choice of noise parameters $\alpha_i$ but to the fortunate form of the right-hand side in the linear system (6).

## 3.1 Proof of Theorem 3.1

*Proof of Theorem 3.1.* By Chebyshev's inequality, the empirical average of $m$ samples of an estimator with variance $v$ is within $2\sqrt{v/m}$ of its mean except with probability $1/4$. By Claim 3.2 and Claim 3.3, with probability at least $3/4$,

$$|\tilde{W}^1[f] - \mathbf{W}^1[f]| \le \|d\|_1 \rho^{\ell-1} \|f^{\ge \ell}\|^2 + \sqrt{\frac{2\rho^{-2}\|d\|_1^2}{m}} \cdot \|f\|_4^2.$$

As $\|f^{\ge \ell}\| \le \|f\|_4$, it is sufficient that the choices of $\rho, \ell, m$ satisfy

$$\|d\|_1 \rho^{\ell-1} \le \frac{\epsilon'}{2} \quad \text{and} \quad \frac{2\rho^{-2}\|d\|_1^2}{m} \le \frac{\epsilon'^2}{4},$$

for $\epsilon' = \epsilon / \|f\|_4^2$. By Claim 3.5, $\|d\|_1 \le \sqrt{\ell} \cdot \|d\|_2 \le \ell^{3/2}$. The choice $\rho = \epsilon'^{(\ell-1)^{-1}}/6$ ensures that for $\ell \ge 2$,

$$\|d\|_1 \rho^{\ell-1} \le \frac{\ell^{3/2}}{6^{\ell-1}} \cdot \epsilon' \le \frac{1}{2}\epsilon'.$$

As $\ell = \log 1/\epsilon' + 1$,

$$\frac{2\rho^{-2}\|d\|_1^2}{m} \le \frac{24\ell^3}{m},$$

which is at most $\epsilon'^2/4$ by our choice $m = 96\ell^3/\epsilon'^2$. □

*Proof of Claim 3.2.* The bias of the estimator is $\mathbf{E}[fg] - \mathbf{W}^1[f] = \mathbf{E}[fg] - \|f^{=1}\|^2$. The first $\ell$ levels of $g$ are given by (3), so $\mathbf{E}[f^{<\ell}g^{<\ell}] = \|f^{=1}\|^2$. As for $j \geq \ell$, by (2),

$$g^{=j} = \rho^{-1} \sum d_i (N_{\alpha_i \rho} f)^{=j} = \sum d_i (\alpha_i \rho)^j f^{=j}$$

from where

$$\begin{aligned}
|\mathbf{E}\,\tilde{W}^1[f] - \mathbf{W}^1[f]| &= |\mathbf{E}[f^{\geq \ell} g^{\geq \ell}]| \\
&\leq \sum_{j \geq \ell} \rho^{j-1} \sum_i |d_i \alpha_i^j| \cdot \|f^{=j}\|^2 && \text{(by triangle inequality)} \\
&\leq \rho^{\ell-1} \|d\|_1 \sum_{j \geq \ell} \|f^{=j}\|^2 && (|\alpha_i| \leq 1) \\
&= \rho^{\ell-1} \|d\|_1 \|f^{\geq \ell}\|^2. && \square
\end{aligned}$$

*Proof of Claim 3.3.* Operator $N_\rho$ is contractive: $\|N_\rho g\| \leq \|g\|$ for all $\rho$ and $g$.

$$\begin{aligned}
\mathbf{Var}\,\tilde{W}^1[f] &\leq \mathbf{E}[\tilde{W}^1[f]^2] \\
&= \rho^{-2} \|d\|_1^2\, \mathbf{E}\,\mathbf{E}[f^2 \cdot N_{\alpha_i \rho}(f^2) | i] \\
&\leq \rho^{-2} \|d\|_1^2\, \mathbf{E} \sqrt{\mathbf{E}[f^4|i] \cdot \mathbf{E}[N_{\alpha_i \rho}(f^2)^2 | i]} && \text{by Cauchy-Schwarz} \\
&\leq \rho^{-2} \|d\|_1^2 \cdot \|f^2\| \cdot \|f^2\| && \text{by contractivity} \\
&= \rho^{-2} \|d\|_1^2 \|f\|_4^4. && \square
\end{aligned}$$

## 3.2 Lower bound

**Theorem 3.8.** *For every $o(1/\epsilon^2)$-query algorithm $A$ there exists a Boolean-valued $f$ such that $|A(f) - \mathbf{W}^1[f]| > \epsilon$ with probability at least $1/4$, as long as $\epsilon = \Omega(n^{3/2} 2^{-n/2})$.*

The advantage of a distinguisher $D$ with respect to random variable $f$ and $g$ is the absolute difference in the probabilities that $D$ accepts $f$ and $g$. The maximum possible distinguishing advantage is known to equal the minimum possible probability of the event $f \neq g$ under all couplings of $f$ and $g$, i.e., joint distributions over $(f, g)$ that are consistent with their marginals.

**Fact 3.9.** Assuming $0 \leq \beta < \alpha < 0.9$, distinguishing between independent $\alpha$ and $\beta$-biased coin flips with advantage at least $1/4$ requires $\Omega(1/(\alpha - \beta)^2)$ samples.

*Proof.* There are several proofs for the case $\beta = 0$, see for example [13]. The general case reduces to this special case. Consider the reduction that takes the outcome of a coin flip, outputs it with probability $1 - \beta$, and outputs 1 otherwise. This reduction maps unbiased coins into $\beta$-biased ones and $(\alpha - \beta)/(1 - \beta)$-biased coins into $\alpha$-biased ones. Distinguishing the latter with advantage $1/4$ therefore requires $\Omega((1 - \beta)^2/(\alpha - \beta)^2)$ samples. $\square$

*Proof of Theorem 3.8.* Choose any $0.5 < \beta < \alpha < 0.9$ with $\alpha - \beta = 5\epsilon$. Let $f$ (resp., $g$) be a probabilistic Boolean function in which the values $f(x)$ are independent $\alpha x_1$-biased, (resp., $\beta x_1$-biased). By Claim 3.10 $\mathbf{W}^1[f]$ is at least $\alpha^2 - \epsilon$, and $\mathbf{W}^1[g]$ is at most $\beta^2 + \epsilon$ except with probability at most $1/4$. Therefore

$$\mathbf{W}^1[f] - \mathbf{W}^1[g] \geq \alpha^2 - \beta^2 - 2\epsilon = (\alpha - \beta)(\alpha + \beta) - 2\epsilon \geq 3\epsilon$$

by our choice of $\alpha$ and $\beta$. Had $A$ been $\epsilon$-accurate with probability $3/4$, by a union bound,

$$\Pr[|A(f) - A(g)| \leq \epsilon] \leq \Pr[|A(f) - \mathbf{W}^1[f]| \geq \epsilon] + \Pr[|A(g) - \mathbf{W}^1[g]| \geq \epsilon] + \Pr[|\mathbf{W}^1[f] - \mathbf{W}^1[g]| \leq 3\epsilon] \leq 3/4$$

so $\Pr[f \neq g] \geq \Pr[|A(f) - A(g)| > \epsilon] > 1/4$. As this holds for an arbitrary coupling of $f$ and $g$, the two can be $1/4$-distinguished.

We now show this is impossible, namely $f$ and $g$ cannot be $1/4$-distinguished with $q = o(1/\epsilon^2)$ queries. For if they could be by some algorithm $A$, then the following algorithm can distinguish $\alpha$ and $\beta$-biased coin flips with $q$ queries, contradicting Fact 3.9: Whenever $A$ makes a new query $x$, flip a coin, multiply the outcome by $x_1$ and provide this answer. The view of $A$ when the reduction provides $\alpha$ and $\beta$-biased coins is identical to interactions with $f$ and $g$, respectively. $\square$

**Claim 3.10.** *Let $h$ be a random Boolean function whose values are independent $\pm 1$ bits. Then $|\mathbf{W}^1[h] - \mathbf{W}^1[\mathbf{E}\, h]| \le \epsilon$ except with probability at most $O(n^3 2^{-n}/\epsilon^2)$.*

*Proof.* By independence, $\mathbf{Var}[\hat{h}(i)] \le 2^{-n}$ for every $i$ so by Chebyshev's inequality, $|\hat{h}(i) - \mathbf{E}\,\hat{h}(i)| \le \epsilon/2n$ except with probability $O(n^2 2^{-n}/\epsilon^2)$. Therefore $|\hat{h}(i)^2 - \mathbf{E}[\hat{h}(i)]^2| = |\hat{h}(i) + \mathbf{E}\,\hat{h}(i)| \cdot |\hat{h}(i) - \mathbf{E}\,\hat{h}(i)| \le \epsilon/n$. The claim follows from a union bound and the triangle inequality. □

For this argument to result in a $\Omega(1/\epsilon^2)$ lower bound it is essential that $\alpha$ and $\beta$ be bounded away both from 0 and from 1: Had they been too close to 1 the coins would have been distinguishable from $O(1/\epsilon)$ samples. Had they been too close to 0, the typical distance between $W^1[f]$ and $W^1[g]$ would have been $O(\epsilon^2)$.

# 4 Learning

In the setting of *distribution-free* (agnostic) learning of a linear approximation from independent random samples, $\Omega(n/\epsilon)$ samples are necessary for properly learning a bounded $n$-dimensional function with respect to expected square loss error [12]. This bound can be matched by an improper learning algorithm but not by empirical risk minimization [14].

Shamir [12] points out that in general, loss minimization (output a linear hypothesis $\ell$ such that $\mathbf{E}[(f(x) - \ell(x))^2]$ is $\epsilon$-close to best possible) and model approximation (find $\ell$ such that $\mathbf{E}[(\ell(x) - \ell^\star(x))^2] \le \epsilon$ for the $\ell^\star$ that minimizes $\mathbf{E}[(f(x) - \ell^\star(x))^2]$) are not equivalent problems.

In the context of learning under the uniform distribution, however, the two problems are equivalent to approximating the projection $f^{\le 1}$ onto the space of linear functions. Linial, Mansour, and Nisan [8] showed that the approximation can be calculated from $O(n/\epsilon)$ samples via empirical risk minimization.

We show a query complexity lower bound of $\Omega(n/\sqrt{\epsilon})$ for proper agnostic learning, even when the learner is given non-adaptive query access to $f$.

**Theorem 4.1.** *For every non-adaptive algorithm $A$ that makes $o(n/\sqrt{\epsilon})$ queries and outputs an affine hypothesis, there is an $f\colon \{\pm 1\}^n \to \{\pm 1\}$ such that $\|f - A(f)\|^2 \le \|f - f^{\le 1}\|^2 + \epsilon$ only with probability $2^{-\Omega(n)}$ assuming $\epsilon \ge 2^{-(1-o(1))n/2}$.*

Our proof technique also gives a $\Omega(\log n/\sqrt{\epsilon})$ query complexity lower bound for adaptive $A$.

Theorem 4.1 is proved in two steps. In Proposition 4.2 we establish it for sufficiently small but constant $\epsilon$. In Proposition 4.5 we give a reduction from agnostically learning linear approximations with constant mean-square error and $q$ queries to the same problem with mean-square error $\epsilon$ and $O(q/\sqrt{\epsilon})$ queries. The reduction is flexible with respect to the learning model, and even to the task. It can also be applied to estimation but gives a worse dependence on $\epsilon$ than what we have in Theorems 3.8 and 5.2.

## 4.1 Lower bound for constant error

**Proposition 4.2.** *For every $\kappa < 1$ there exist $\epsilon, \eta$ such that for $n$ sufficiently large and for every non-adaptive algorithm $A$ that makes at most $\kappa n$ queries and outputs an affine hypothesis, there exists $f\colon \{\pm 1\}^n \to \{\pm 1\}$ such that $\|f - A(f)\|^2 \le \|f - f^{\le 1}\|^2 + \epsilon$ with probability at most $2^{-\eta n}$.*

A $\log n - O(1)$ adaptive query complexity lower bound follows from the same assumptions as adaptive algorithms that make $q$ Boolean-valued queries can be simulated with $2^q$ non-adaptive queries.

The proof relies on the following lemma, which states that there is a large set of Boolean functions whose close-to-best linear approximations are mutually far apart. Any candidate learning algorithm with too few queries does not have enough information to disambiguate between these functions and is unlikely to output an approximately correct hypothesis.

**Lemma 4.3.** *For every $\rho < 1$ there exists $\epsilon > 0$ such that for sufficiently large $n$, there is a set $\mathcal{P}$ of $2^{\rho n}$ pairs of functions $(f, \ell)$ over domain $\{-1, 1\}^n$, where $f$ is boolean-valued, $\ell$ is linear and real-valued, and*

1. *For every $(f, \ell) \in \mathcal{P}$, $\|\ell - f^{\le 1}\| \le \sqrt{\epsilon}$.*
2. *For all distinct $(f, \ell), (f', \ell') \in \mathcal{P}$, $\|\ell - \ell'\| \ge 4\sqrt{\epsilon}$.*

By the triangle inequality:

**Corollary 4.4.** *For every $\rho < 1$ there exists $\epsilon > 0$ and a set $\mathcal{F}$ of $2^{\rho n}$ Boolean functions such that for every $f \neq g \in \mathcal{F}$, $f^{\leq 1}$ and $g^{\leq 1}$ are at least $2\sqrt{\epsilon}$ far apart.*

By Yao's Minimax principle, to prove proposition 4.2 it suffices (and is necessary) to give a probability distribution over the possible inputs $f$ and show that any non-adaptive algorithm that makes too few queries is likely to fail on this distribution. Our distribution is uniform over the collection from Corollary 4.4: As $\log|\mathcal{F}|$ is larger than the query complexity, the algorithm does not have enough information to disambiguate between candidate inputs in $\mathcal{F}$ and is therefore unlikely to be correct.

*Proof of proposition 4.2.* Choose $f$ at random from the set of functions $\mathcal{F}$ in Corollary 4.4 instantiated with $\rho$ satisfying $1-\rho = (1-\kappa)/2$. Let $f_Q$ denote the restriction of $f$ on the query set $Q$. Conditioned on the choice of $Q$, by the law of total expectation

$$\begin{aligned}
\mathbf{E}\,\frac{1}{|\{g \in \mathcal{F}\colon g_Q = f_Q\}|} &= \sum_{a \in \mathrm{Supp} f_Q} \frac{1}{|\{g \in \mathcal{F}\colon g_Q = a\}|} \cdot \Pr_{f\sim\mathcal{F}}[f_Q = a] \\
&= \sum_{a \in \mathrm{Supp} f_Q} \frac{1}{|\mathcal{F}| \Pr_{g\sim\mathcal{F}}[g_Q = a]} \cdot \Pr_{f\sim\mathcal{F}}[f_Q = a] \\
&= \frac{|\mathrm{Supp} f_Q|}{|\mathcal{F}|} \\
&\leq \frac{2^q}{|\mathcal{F}|},
\end{aligned}$$

where $q$ is the query complexity. By our choice of parameters this is at most $2^{-(1-\kappa)n/2}$.

The left-hand side upper bounds the probability that the algorithm succeeds. For conditioned on the view of $A$ and the choice of $f$, its input is equally likely to have been any function in the set $S_f = \{g \in \mathcal{F}\colon g_Q = f_Q\}$. However, the output $h$ of $A(f)$ could be accurate for at most one function in this set: If

$$\begin{aligned}
\|g - h\|^2 &\leq \|g^{\leq 1} - g\|^2 + \epsilon \quad \text{and} \\
\|g' - h\|^2 &\leq \|g'^{\leq 1} - g'\|^2 + \epsilon
\end{aligned}$$

by Pythagoras' theorem $\|g^{\leq 1} - h\|^2, \|g'^{\leq 1} - h\|^2 \leq \epsilon$. By the triangle inequality $g^{\leq 1}$ and $g'^{\leq 1}$ are $2\sqrt{\epsilon}$-close, so it must be that $g = g'$. Therefore $A(f)$ succeeds with probability at most $1/|S_f|$. □

## 4.2 Trading accuracy for queries

We describe a reduction $R$ that, given access to a high-accuracy learner $A$ that requires access to many queries, learns with low accuracy but fewer queries.

When $A$ makes a query, the reduction flips a coin with success probability $\delta = 2\sqrt{\epsilon/\epsilon_0}$. If the coin flip succeeds, the reduction forwards the query and returns the answer to $A$. If it fails, the reduction answers the query randomly. When $A$ outputs its answer $h$, the reduction returns $h/\delta$.

The reduction can be implemented in any query model (random samples, non-adaptive, adaptive). We show that it is effective in the context of approximately learning linear approximations.

We show correctness for Boolean-valued functions as it matches our application, but the proposition should hold more generally under any $p$-norm bound.

**Proposition 4.5.** *If $A$ makes $q$ queries and outputs $h$ such that $\|h - f^{\leq 1}\|^2 \leq \epsilon$ given Boolean-valued $f$ as input, then $R^A(f)$ makes at most $q' = 4\sqrt{\epsilon/\epsilon_0}q$ queries except with probability $2^{-\Omega(q')}$, and outputs $h'$ such that $\|h' - f^{\leq 1}\|^2 \leq \epsilon_0$ except with probability $O(n/\epsilon 2^n)$.*

*Proof. Query complexity:* The number of queries is a Binomial$(q, \delta)$ random variable. By a Chernoff bound it exceeds $2\delta q$ with probability at most $2^{-\Omega(\delta q)}$.

*Correctness:* The input to $A$ provided by $R$ is indistinguishable from a function $f'$ obtained by corrupting $f$ by noise of rate $\delta$, namely $f'(x) = N(x)f(x)$, where $N(x)$ are independent $\delta$-biased bits. By the triangle inequality,

$$\|h/\delta - f^{\leq 1}\| \leq \frac{1}{\delta}\|h - f'^{\leq 1}\| + \frac{1}{\delta}\|f'^{\leq 1} - \delta f^{\leq 1}\|.$$

By our choice of $\delta$, the first term on the right is at most $\sqrt{\epsilon_0}/2$, so it suffices to upper bound the second one by that much. In expectation, using Parseval's identity,

$$\mathbf{E}\|f'^{\leq 1} - \delta f^{\leq 1}\|^2 = \frac{(1-\delta^2)(n+1)}{2^n}, \tag{8}$$

By Markov's inequality, $\delta^{-1}\|f'^{\leq 1} - \delta f^{\leq 1}\| \leq \sqrt{\epsilon_0}/2$ except with probability $O(n/\epsilon 2^n)$. □

As an aside, $R$ can be implemented in a "complexity-preserving" manner in the sense that if $A$ is only guaranteed to work on "low-complexity" inputs $f$ (e.g. those of small circuit complexity) then $R^A$ works on all inputs of slightly lower complexity. The reason is that Proposition 4.5 (specifically (8)) only relies on the *pairwise* independence of the noise.

*Proof of Theorem 4.1.* Fix $\kappa = 1/2$ and let $\epsilon_0$ be the corresponding $\epsilon$ from Proposition 4.2. Suppose $A$ succeeds with higher probability. By Proposition 4.5 $R^A$ makes $o(n) < n/2$ queries, succeeds with probability $2^{-o(n)}$, and outputs an $\epsilon_0$-approximation, violating Proposition 4.2. □

A $\Omega(\log n/\sqrt{\epsilon})$ adaptive query lower bound follows from the same argument.

## 4.3 Proof of Lemma 4.3

**Claim 4.6.** *For every $\epsilon$ there exists a Boolean function $f\colon \{-1,1\}^n \to \{-1,1\}$ such that $\|f^{\leq 1} - \ell\|^2 \leq \epsilon$ for sufficiently large $n$, where $\ell(x) = (\gamma/\sqrt{n})(x_1 + \cdots + x_n)$, assuming $4\sqrt{6}\gamma^2 \exp(-1/4\gamma^2) \leq \epsilon$.*

The proof uses the following special case of Khintchine's inequality:

**Fact 4.7** (Khintchine's inequality)**.** For every linear function $\ell : \{-1,1\}^n \to \mathbb{R}$, $\|\ell\|_4^4 \leq 3\|\ell\|_2^4$.

*Proof of Claim 4.6.* We show that a random $f$ from a suitable distribution satisfies the desired property with nonzero probability. For each $x$, $f(x)$ is a random $\{-1,1\}$ outcome with bias

$$\ell(x)\mathbb{1}(|\ell(x)| \leq 1) = \begin{cases} \ell(x), & \text{if } |\ell(x)| \leq 1 \\ 0, & \text{if not.} \end{cases}$$

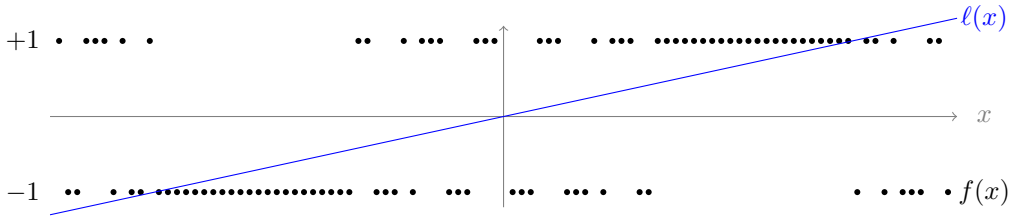


Figure 1: Illustration of $f$. When $|\ell(x)| \leq 1$, $f(x)$ is $\ell(x)$-biased. Otherwise, it is unbiased.

The values of $f$ are chosen independently conditioned on $f(-x) = -f(x)$ for every $x$. (This folding simplifies the proof a little bit.) The construction is illustrated in Figure 1. By construction, $f$ is balanced. By Parseval's identity,

$$\|f^{\leq 1} - \ell\|^2 = \sum_{i=1}^{n} (\hat{f}(i) - \gamma/\sqrt{n})^2. \tag{9}$$

We analyze the concentration of $\hat{f}(i)$. In expectation, and by linearity of expectation,

$$\begin{aligned}\mathbf{E}\,\hat{f}(i) &= \mathbf{E}[x_i f(x)] \\ &= \mathbf{E}\,\mathbf{E}[x_i f(x)|x] \\ &= \mathbf{E}[x_i \ell(x) \mathbb{1}(|\ell(x)| \le 1)] \\ &= \mathbf{E}[x_i \ell(x)] - \mathbf{E}[x_i \ell(x) \mathbb{1}(|\ell(x)| > 1)] \\ &= \frac{\gamma}{\sqrt{n}} - \hat{g}(i),\end{aligned}$$

where $g$ is the function $g(x) = \ell(x)\mathbb{1}(|\ell(x)| > 1)$. As $g$ is a symmetric function, its first-level Fourier coefficients are equal and must therefore have absolute value

$$\begin{aligned}|\hat{g}(i)| &= \sqrt{\frac{\hat{g}(1)^2 + \cdots + \hat{g}(n)^2}{n}} \\ &\le \sqrt{\frac{\mathbf{E}[g(x)^2]}{n}} && \text{by Parseval's identity} \\ &= \frac{1}{\sqrt{n}} \sqrt{\mathbf{E}[\ell(x)^2 \mathbb{1}(|\ell(x)| > 1)]} \\ &\le \frac{1}{\sqrt{n}} \sqrt[4]{\mathbf{E}[\ell(x)^4] \Pr[|\ell(x)| > 1]} && \text{by Cauchy-Schwarz} \\ &\le \frac{1}{\sqrt{n}} \sqrt[4]{3\|\ell\|^4 \Pr[|\ell(x)| > 1]} && \text{by Fact 4.7} \\ &\le \frac{\sqrt[4]{6}\gamma}{\sqrt{n}} \exp\left(-\frac{1}{8\gamma^2}\right) && \text{by Hoeffding's bound} \\ &\le \sqrt{\epsilon/4n} && \text{by the assumption on } \gamma.\end{aligned}$$

As $\hat{f}(i)$ is the average of $2^n/2$ independent outcomes, by Hoeffding's bound

$$\Pr\left[|\hat{f}(i) - \mathbf{E}\,\hat{f}(i)| > \sqrt{\epsilon/4n}\right] \le 2\exp(-2^n\epsilon/8n).$$

Assuming $n$ is sufficiently large so that $2\exp(-2^n\epsilon/8n)$ is less than $1/n$, by the triangle inequality and a union bound, $|\hat{f}(i) - \gamma/\sqrt{n}| \le \sqrt{\epsilon/n}$ for all $i$ with positive probability. By (9) there must exist a choice of $f$ for which $\|f^{\le 1} - \ell\|^2 \le \epsilon$. □

*Proof of lemma 4.3.* Let $\gamma$ satisfy $4\sqrt{6}\gamma^2\exp(-1/4\gamma^2) = \epsilon$, $\mathcal{C}$ be a code over $\{-1,1\}^n$ of rate $\rho$ and mininum relative Hamming distance $4\epsilon/\gamma^2$. Such codes exist for sufficiently large $n$ by the Gilbert-Varshamov bound. Let $(f,\ell)$ be the pair of functions from Claim 4.6. The collection $\mathcal{P}$ consists of the pairs $(f_\sigma, \ell_\sigma)\colon \sigma \in \mathcal{C}$, where $g_\sigma(x)$ is the function $g(\sigma_1 x_1, \ldots, \sigma_n x_n)$.

As $g_\sigma$ is a permutation of $g$, $\|f_\sigma^{\le 1} - \ell_\sigma^{\le 1}\|^2 = \|f - \ell\|^2 \le \epsilon$ proving property 1. As for property 2, by Parseval's identity, for $\sigma \ne \sigma' \in \mathcal{C}$,

$$\|\ell_\sigma - \ell_{\sigma'}\|^2 = \sum \frac{\gamma^2}{n}(\sigma_i - \sigma_i')^2 = \frac{4\gamma^2}{n}\sum \mathbb{1}(\sigma_i \ne \sigma_i') \ge 16\epsilon$$

by our choice of parameters. □

# 5 Estimation from uniform samples

The queries in Algorithm 1 come in pairwise correlated pairs $(x, x \cdot e_i)$. They can be emulated from independent pairs $(x,y)$ after reweighting by the likelihood ratio thereby preserving the bias of the estimator. However, the variance of the likelihood ratio grows exponentially in $\Theta(n\rho^2)$ leading to a comparable blowup in the estimator variance (which becomes $\approx 2^{\Omega(n\rho^2)}/\rho^2$). Such an algorithm is not query efficient as the seemingly optimal choice of $\rho \approx 1/\sqrt{n}$ still fails to attain sublinear complexity.

## 5.1 The algorithm

Instead of adapting Algorithm 1, our pair-based sample mean estimator directly calculates the empirical mean of the unbiased estimator $A(x,y) = f(x)f(y)\langle x,y\rangle$ averaged over all pairs of samples. The reuse of samples creates correlations, but their covariances are dominated by their individual variances.

**Algorithm 2** pair-based sample mean estimator

given samples $x^1, \dots x^m, \overset{i.i.d}{\sim} \{-1,1\}^n$
**return** $\tilde{W}^1[f] = \binom{m}{2}^{-1} \sum_{i<j}^{m} A(x^i, x^j)$, where $A(x,y) = f(x)f(y)\langle x,y\rangle$

The algorithm can be implemented in complexity linear in $mn$ (the bit complexity of the samples) by evaluating the equivalent formula

$$\tilde{W}^1[f] = \binom{m}{2}^{-1} \cdot \left( \sum_i \left( \sum\nolimits_j f(x^j) x_i^j \right)^2 - n \sum_j f(x^j)^2 \right).$$

**Theorem 5.1.** *For $m = O(\sqrt{n}\|f\|_4^2/\epsilon + \|f\|_4^4/\epsilon^2)$ independent and uniform samples, $|\tilde{W}^1[f] - \mathbf{W}^1[f]| \le \epsilon$ with probability at least $2/3$.*

*Proof.* When $x$ and $y$ are random, $A$ and therefore $\tilde{W}^1[f]$ is an unbiased estimator of $\mathbf{W}^1[f]$:

$$\begin{aligned}
\mathbf{E}\left[A(x,y)\right] &= \mathbf{E}\left[f(x)f(y)\langle x,y\rangle\right] \\
&= \sum_{i=1}^{n} \mathbf{E}\left[f(x)x_i f(y)y_i\right] \\
&= \sum_{i=1}^{n} \mathbf{E}\left[f(x)x_i\right]^2 \\
&= \sum_{i=1}^{n} \hat{f}(i)^2 = \mathbf{W}^1[f]
\end{aligned}$$

The variance of $\tilde{W}^1[f]$ is the sum of the (co)variances of pairs $A(x^i,x^j)$ and $A(x^{i'},x^{j'})$ indexed by $i<j$ and $i'<j'$ scaled by $\binom{m}{2}^{-2}$. Only intersecting pairs have nonzero contribution. There are $\binom{m}{2}$ and $2(m-2)\binom{m}{2}$ pairs of intersection size 2 and 1, respectively, resulting in

$$\mathbf{Var}\,\tilde{W}^1[f] = \frac{2}{m(m-1)} v + \frac{4(m-2)}{m(m-1)} c$$

where for independent $x, y, z$,

$$\begin{aligned}
v &= \mathbf{Var}\,A(x,y) \\
&\le \mathbf{E}\,A(x,y)^2 \\
&= \mathbf{E}\,f(x)^2 f(y)^2 \langle x,y\rangle^2 \\
&\le \sqrt{\mathbf{E}\,f(x)^4 f(y)^4}\sqrt{\mathbf{E}\langle x,y\rangle^4} && \text{by Cauchy-Schwarz} \\
&= \|f\|_4^4 \cdot \sqrt{3n} && \text{by Fact 4.7}
\end{aligned}$$

and

$$\begin{aligned}
c &= \mathbf{Cov}[A(x,y), A(x,z)] \\
&\le \mathbf{E}\, A(x,y)A(x,z) \\
&= \mathbf{E}\,\mathbf{E}[A(x,y) \mid x]^2 \\
&= \mathbf{E}\left[f(x)^2\,\mathbf{E}\left[f(y)\langle x,y\rangle \mid x\right]^2\right] \\
&\le \sqrt{\|f\|_4^4\,\mathbf{E}\,\mathbf{E}\left[f(y)\langle x,y\rangle \mid x\right]^4} && \text{by Cauchy-Schwarz} \\
&= \|f\|_4^2 \cdot \|f^{=1}\|_4^2 && \text{by Plancherel's formula} \\
&= \|f\|_4^2 \cdot \sqrt{3}\|f^{=1}\|_2^2 && \text{by Fact 4.7} \\
&\le \sqrt{3}\|f\|_4^4.
\end{aligned}$$

Summing up, $\mathbf{Var}\,\tilde{W}^1[f] = O(n/m^2 + 1/m) \cdot \|f\|_4^4$. The conclusion follows by applying Chebyshev's inequality to $\tilde{W}^1[f]$. □

## 5.2 Lower bound

We show that for sample-based algorithms (unlike query-based) for this problem the dependence on the dimension is inherent. More precisely, the following theorem say that Algorithm 2 is optimal in the sample model.

**Theorem 5.2.** *For every (possibly randomized) $A$ there exists $f\colon \{\pm1\}^n \to \{\pm1\}$ so that when given $m$ independent examples $(x, f(x))$ with uniform $x$, $A$ outputs a value within $\epsilon$ of $\mathbf{W}^1[f]$ with probability at most $1/2 + O(m\epsilon/\sqrt{n}) + O((n/\epsilon)2^{-n})$.*

The proof in fact shows that the query complexity bound in Theorem 5.1 is tight in all parameters, even if $\|f\|_4$ is weakened to $\|f\|_\infty$ in the statement.

To prove Theorem 5.2, we construct a function $f$ that is weakly pseudorandom given $o(\sqrt{n}/\epsilon)$ examples but for which $\mathbf{W}^1[f]$ is likely to be $\Omega(\epsilon)$. In contrast, $\mathbf{W}^1[f]$ for a random function is concentrated around $2^{-n}$.

The function $f$ is the sign of

$$g_{Z,N}(x) = \sqrt{\epsilon/n} \cdot \langle x, Z\rangle + \sqrt{1-\epsilon} \cdot N(x),$$

where $Z$ is an $n$-dimensional standard normal and $N$ is a standard normal function over $\{-1,1\}^n$ (the $2^n$ values $N(x)$ are independent standard normals as $x$ ranges over $\{-1,1\}^n$.)

**Claim 5.3.** *The statistical distance between $(x^1, O(x^1)), \ldots, (x^m, O(x^m))$ when $O = f$ and $O$ is a random function is at most $O(m\epsilon/\sqrt{n})$.*

*Proof.* It is sufficient to show the claim for the real-valued functions $g = g_{Z,N}$ and $N$, as $f$ and a random Boolean function are obtained by taking signs of $g$ and $N$, respectively, and postprocessing cannot increase statistical distance.

We will assume without loss of generality that $x^1, \ldots, x^m$ are all distinct as repetitions can only decrease statistical distance. Fixing $x^1, \ldots, x^m$, $g(x^1), \ldots, g(x^m)$ are jointly centered Gaussian with covariance matrix

$$\Sigma_{ij} = \begin{cases} (\epsilon/n)\langle x^i, x^j\rangle & \text{if } i \ne j, \\ 1, & \text{if } i = j \end{cases}$$

The covariance matrix of $N(x^1), \ldots, N(x^m)$ is the identity $Id$. The conditional statistical distance given $x^1, \ldots, x^m$ is at most [4]

$$O(\|\Sigma - Id\|_F) = O\left(\frac{\epsilon}{n}\sqrt{\textstyle\sum_{i\ne j}\langle x^i, x^j\rangle^2}\right).$$

By Cauchy-Schwarz, the average, unconditional statistical distance is at most

$$O\left(\frac{\epsilon}{n}\sqrt{\textstyle\sum_{i\ne j}\mathbf{E}\langle x^i, x^j\rangle^2}\right) = O\left(\frac{\epsilon}{n}\sqrt{m^2 \cdot n}\right) = O(m\epsilon/\sqrt{n}).$$

□

**Claim 5.4.** *For every $z$ such that $n/64 \le \|z\|^2 \le 4n$, $\mathbf{E}[\langle X, z\rangle \operatorname{sign} g_{z,N}(X)|N] = \Omega(\sqrt{\epsilon n})$ except with probability $O(\epsilon \cdot 2^{-n})$.*

*Proof.* The expression inside the expectation is (up to a $\sqrt{n}$ factor) of the form $t \operatorname{sign}(\sqrt{\epsilon}t + \sqrt{1-\epsilon}N)$ for $t = \langle X, z\rangle/\sqrt{n}$. For fixed $t$ and standard normal $N$

$$\mathbf{E}[t \operatorname{sign}(\sqrt{\epsilon}t + \sqrt{1-\epsilon}N)] = 2|t| \Pr(|N| < \sqrt{\epsilon/(1-\epsilon)}|t|) \ge 2|t| \Pr(|N| < \sqrt{\epsilon}|t|). \tag{10}$$

As the right-hand side is always nonnegative,

$$\begin{aligned}
&\mathbf{E}[\langle X, z\rangle \operatorname{sign} g_{z,N}(X)] \\
&\quad \ge \mathbf{E}[\langle X, z\rangle \operatorname{sign} g_{z,N}(X) | |\langle X, z\rangle| \ge \|z\|/3] \cdot \Pr(|\langle X, z\rangle| \ge \|z\|/3) \\
&\quad \ge 2(\|z\|/3) \Pr(|N| < \sqrt{\epsilon}/24) \cdot \Pr(|\langle X, z\rangle| \ge \|z\|/3) && \text{by (10)} \\
&\quad = \Omega(\sqrt{n} \cdot \Pr(|N| < \sqrt{\epsilon}/24 \cdot 3)) && \text{by Khinchine's inequality} \\
&\quad = \Omega(\sqrt{\epsilon n}).
\end{aligned}$$

By independence of the values $\operatorname{sign} g_{z,N}(x)$ across $x$,

$$\begin{aligned}
\mathbf{Var}\, \mathbf{E}[\langle X, z\rangle \operatorname{sign} g_{z,N}(X)|N] &= \mathbf{Var}\, \frac{1}{2^n} \sum_x \langle x, z\rangle \operatorname{sign} g_{z,N}(x) \\
&= \frac{1}{2^{2n}} \sum_x \langle x, z\rangle^2\, \mathbf{Var} \operatorname{sign} g_{z,N}(x) \\
&\le \frac{1}{2^{2n}} \sum_x \langle x, z\rangle^2 \\
&= \frac{\|z\|^2}{2^n}.
\end{aligned}$$

The claim follows from Chebyshev's inequality. □

**Claim 5.5.** $\mathbf{W}^1[f] \ge \Omega(\epsilon)$ *except with probability* $\Omega(2^{-n})$.

*Proof.* By the Cauchy-Schwarz inequality, for any linear function $\ell$,

$$\langle f, \ell\rangle^2 = \langle f^{\le 1}, \ell\rangle^2 \le \|f^{\le 1}\|^2 \cdot \|\ell\|^2 = \mathbf{W}^1[f] \cdot \|\ell\|^2.$$

The function $\ell(x) = \langle x, Z\rangle$ has squared 2-norm $\sum Z_i^2$, which is between $n/64$ and $4n$ except with probability $2^{-n}$. Applying Claim 5.4,

$$\mathbf{W}^1[f] \ge \frac{\langle f, \ell\rangle^2}{\|\ell\|^2} = \frac{\mathbf{E}[\langle X, Z\rangle \operatorname{sign} g_{Z,N}(X)|N, Z]^2}{\mathbf{E}[\langle X, Z\rangle^2|Z]} = \Omega(\epsilon).$$ □

**Claim 5.6.** *For a random Boolean function $r$, $\mathbf{W}^1[r] \le \epsilon$ except with probability $(n/\epsilon)2^{-n}$.*

*Proof.* By symmetry the expectation is $n2^{-n}$. The bound follows from Markov's inequality. □

*Proof of Theorem 5.2.* Applying Claims 5.5 with $\epsilon$ in the definition of $f$ scaled up by a suitable constant factor, $\mathbf{W}^1[f] > 3\epsilon$ except with probability $O(2^{-n})$. By Claim 5.6, $\mathbf{W}^1[r] \le \epsilon$ except with probability $(n/\epsilon)2^{-n}$.

Let $D$ be the distinguisher that accepts if $A$'s output is greater than $2\epsilon$ and rejects if not. If $A$ is correct with probability at least $(1+\delta)/2$ on every input, then $D$ accepts $f$ with probability at least $1/2 + \delta/2 - O(2^{-n})$ while $D$ accepts $r$ with probability at most $1/2 - \delta/2 + (n/\epsilon)2^{-n}$. As $D$'s distinguishing advantage cannot exceed $A$'s, the statistical distance between $A$'s views on inputs $f$ and $r$ is at least $\delta - O((n/\epsilon)2^{-n})$. By Claim 5.3 it is at most $O(m\epsilon/\sqrt{n})$, from where $\delta \le O(m\epsilon/\sqrt{n}) + O((n/\epsilon)2^{-n})$. □

## Acknowledgments

This work was supported by an NSERC discovery grant. We thank Gautam Prakriya for insightful discussions and the anonymous ITCS reviewers for corrections and helpful comments.

# References


[1] Mitali Bafna, Srikanth Srinivasan, and Madhu Sudan. Local decoding and testing of polynomials over grids. *Random Struct. Algorithms*, 57(3):658–694, 2020.

[2] Shai Ben-David and Ruth Urner. The sample complexity of agnostic learning under deterministic labels. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 527–542, 13–15 Jun 2014.

[3] Andrej Bogdanov and Gautam Prakriya. Direct Sum and Partitionability Testing over General Groups. In *48th International Colloquium on Automata, Languages, and Programming (ICALP 2021)*, volume 198, pages 33:1–33:19, 2021.

[4] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

[5] Noah Fleming and Yuichi Yoshida. Distribution-free testing of linear functions on $\mathbb{R}^n$. *arXiv preprint arXiv:1909.03391*, 2019.

[6] Shirley Halevy and Eyal Kushilevitz. Distribution-free property-testing. *SIAM J. Comput.*, 37(4):1107–1138, 2007.

[7] Subhash Khot and Dana Moshkovitz. NP-hardness of approximately solving linear equations over reals. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 413–420, 2011.

[8] Nathan Linial, Yishay Mansour, and Noam Nisan. Constant depth circuits, fourier transform, and learnability. *Journal of the ACM (JACM)*, 40(3):607–620, 1993.

[9] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[10] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *J. Comput. Syst. Sci.*, 72(6):1012–1042, 2006.

[11] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014.

[12] Ohad Shamir. The sample complexity of learning linear predictors with the squared loss. *J. Mach. Learn. Res.*, 16(1):3475–3486, jan 2015.

[13] Madhur Tulsiani. Lecture 5: Information and coding theory. Lecture Notes, Winter 2021. `https://home.ttic.edu/~madhurt/courses/infotheory2021/l5.pdf`.

[14] Tomas Vaškevičius and Nikita Zhivotovskiy. Suboptimality of constrained least squares and improvements via non-linear predictors. *Bernoulli*, 29, 02 2023.