# Active Regression via Linear-Sample Sparsification


Xue Chen*
xue.chen1@northwestern.edu
Northwestern University

Eric Price†
ecprice@cs.utexas.edu
The University of Texas at Austin


March 15, 2019


**Abstract**

We present an approach that improves the sample complexity for a variety of curve fitting problems, including active learning for linear regression, polynomial regression, and continuous sparse Fourier transforms. In the active linear regression problem, one would like to estimate the least squares solution $\beta^*$ minimizing $\|X\beta - y\|_2$ given the entire unlabeled dataset $X \in \mathbb{R}^{n\times d}$ but only observing a small number of labels $y_i$. We show that $O(d/\varepsilon)$ labels suffice to find a $\varepsilon$-approximation $\widetilde{\beta}$ to $\beta^*$:

$$\mathbb{E}[\|X\widetilde{\beta} - X\beta^*\|_2^2] \le \varepsilon \|X\beta^* - y\|_2^2.$$

This improves on the best previous result of $O(d \log d + d/\varepsilon)$ from leverage score sampling. We also present results for the *inductive* setting, showing when $\widetilde{\beta}$ will generalize to fresh samples; these apply to continuous settings such as polynomial regression. Finally, we show how the techniques yield improved results for the non-linear sparse Fourier transform setting.



*Supported by NSF Grant CCF-1526952 and a Simons Investigator Award (#409864, David Zuckerman), part of this work was done while the author was a student in University of Texas at Austin and visiting the Simons Institute for the Theory of Computing.

†This work was done in part while the author was visiting the Simons Institute for the Theory of Computing.

# 1 Introduction

We consider the query complexity of recovering a signal $f(x)$ in a given family $\mathcal{F}$ from noisy observations. This problem takes many forms depending on the family $\mathcal{F}$, the access model, the desired approximation norms, and the measurement distribution. In this work, we consider the $\ell_2$ norm and use $D$ to denote the distribution on the domain of $\mathcal{F}$ measuring the distance between different functions, which is not necessarily known to our algorithms.

Our main results are sampling mechanisms that improve the query complexity and guarantees for two specific families of functions — linear families and continuous sparse Fourier transforms.

**Active Linear Regression on a Finite Domain.** We start with the classical problem of linear regression, which involves a matrix $X \in \mathbb{R}^{n \times d}$ representing $n$ points with $d$ features, and a vector $y \in \mathbb{R}^n$ representing the labels associated with those points. The least squares ERM is

$$\beta^* := \arg\min \|X\beta - y\|_2^2.$$

In one *active learning* setting, we receive the entire matrix $X$ but not the entire set of labels $y$ (e.g., receiving any given $y_i$ requires paying someone to label it). Instead, we can pick a small subset $S \subseteq [n]$ of size $m \ll n$, observe $y_S$, and must output $\widetilde{\beta}$ that accurately predicts the entire set of labels $y$. In particular, one would like

$$\|X\widetilde{\beta} - y\|_2^2 \leq (1+\varepsilon)\|X\beta^* - y\|_2^2$$

or (equivalently, up to constants in $\varepsilon$)

$$\|X\widetilde{\beta} - X\beta^*\|_2^2 \leq \varepsilon\|X\beta^* - y\|_2^2. \tag{1}$$

This is known as the "transductive" setting, because it only considers the prediction error on the given set of points $X$; in the next section we will consider the "inductive" setting where the sample points $X_i$ are drawn from an unknown distribution and we care about the generalization to fresh points.

The simplest approach to achieve (1) would be to sample $S$ uniformly over $[n]$. However, depending on the matrix, the resulting query complexity $m$ can be very large – for example, if one row is orthogonal to all the others, it must be sampled to succeed, making $m \geq n$ for this approach.

A long line of research has studied how to improve the query complexity by adopting some form of importance sampling. Most notably, sampling proportional to the leverage scores of the matrix $X$ improves the sample complexity to $O(d \log d + d/\varepsilon)$ (see, e.g., [Mah11]).

In this work, we give an algorithm that improves this to $O(d/\varepsilon)$, which we show is optimal. The $O(d \log d)$ term in leverage score sampling comes from the coupon-collector problem, which is inherent to any i.i.d. sampling procedure. By using the randomized linear-sample spectral sparsification algorithm of Lee and Sun [LS15], we can avoid this term. Note that not every linear spectral sparsifier would suffice for our purposes: deterministic algorithms like [BSS12] cannot achieve (1) for $m \ll n$. We exploit the particular behavior of [LS15] to bound the expected noise in each step.

**Theorem 1.1.** *Given any $n \times d$ matrix $X$ and vector $\vec{y} \in \mathbb{R}^n$, let $\beta^* = \underset{\beta \in \mathbb{R}^d}{\arg\min}\|X\beta - \vec{y}\|_2^2$. For any $\varepsilon < 1$, we present an efficient randomized algorithm that looks at $X$ and produces a diagonal matrix $W_S$ with support $S \subseteq [n]$ of size $|S| \leq O(d/\varepsilon)$, such that*

$$\widetilde{\beta} := \underset{\beta}{\arg\min}\|W_S X \cdot \beta - W_S \cdot \vec{y}\|_2$$

*satisfies*

$$\mathbb{E}\left[\|X \cdot \widetilde{\beta} - X \cdot \beta^*\|_2^2\right] \le \varepsilon \cdot \|X \cdot \beta^* - \vec{y}\|_2^2.$$

*In particular, this implies* $\|X \cdot \widetilde{\beta} - \vec{y}\|_2 \le \big(1 + O(\varepsilon)\big) \cdot \|X \cdot \beta^* - \vec{y}\|_2$ *with 99% probability.*

At the same time, we provide a theoretic information lower bound $m = \Omega(d/\varepsilon)$ matching the query complexity up to a constant factor, when $\vec{y}$ is $X\beta^*$ plus i.i.d. Gaussian noise.

**Generalization for Active Linear Regression.** We now consider the inductive setting, where the $(x, y)$ pairs come from some unknown distribution over $\mathbb{R}^d \times \mathbb{R}$. As in the transductive setting, we see $n$ unlabeled points $X \in \mathbb{R}^{n \times d}$, choose a subset $S \subset [n]$ of size $m$ to receive the labels $y_S$ for, and output $\widetilde{\beta}$. However, the guarantee we want is now with respect to the unknown distribution: for

$$\beta^* := \arg\min \mathop{\mathbb{E}}_{x,y}[(x^T\beta - y)^2],$$

we would like

$$\mathop{\mathbb{E}}_{x,y}[(x^T\widetilde{\beta} - y)^2] \le (1 + \varepsilon) \mathop{\mathbb{E}}_{x,y}[(x^T\beta^* - y)^2]$$

or (equivalently, up to constants in $\varepsilon$)

$$\mathop{\mathbb{E}}_{x}[(x^T\widetilde{\beta} - x^T\beta^*)^2] \le \varepsilon \mathop{\mathbb{E}}_{x,y}[(x^T\beta^* - y)^2].$$

In this inductive setting, there are now two parameters we would like to optimize: the number of labels $m$ and the number of unlabeled points $n$. Our main result shows that there is no significant tradeoff between the two: as soon $n$ is large enough that the ERM for a fully labeled dataset would generalize well, one can apply Theorem 1.1 to only label $O(d/\varepsilon)$ points; and even with an infinitely large unlabeled data set, one would still require $\Theta(d/\varepsilon)$ labels.

But how many unlabeled points do we need for the ERM to generalize? To study this, we consider a change in notation that makes it more natural to consider problems like polynomial regression. In polynomial regression, suppose that $y \approx p(x)$, for $p$ a degree $d-1$ polynomial and $x$ on $[-1, 1]$. This is just a change in notation, since one could express $p(x)$ as $(1, x, ..., x^{d-1})^T\beta$ for some $\beta$. How many observations $y_i = p(x_i) + g(x_i)$ do we need to learn the polynomial, in the sense that

$$\mathop{\mathbb{E}}_{x \in [-1,1]}[(\widetilde{p}(x) - p(x))^2] \le O(1) \cdot \mathbb{E}[g(x)^2]?$$

If we sample $x$ uniformly on $[-1, 1]$, then $O(d^2)$ samples are necessary; if we sample $x$ proportional to $\frac{1}{\sqrt{1-x^2}}$, then $O(d \log d)$ samples suffice (this is effectively leverage score sampling); and if we sample $x$ more carefully, we can bring this down to $O(d)$ [CDL13, CKPS16]. This work shows how to perform similarly for *any* linear family of functions, including multivariate polynomials. We also extend the result to *unknown* distributions on $x$.

In the model we consider, then, $x$ is drawn from an unknown distribution $D$ over an arbitrary domain $G$, and $y = y(x_i)$ is sampled from another unknown distribution conditioned on $x_i$. We are given a dimension-$d$ linear family $\mathcal{F}$ of functions $f : G \to \mathbb{C}$. Given $n$ samples $x_i$, we can pick $m$ of the $y_i$ to observe, and would like to output a hypothesis $\widetilde{f} \in \mathcal{F}$ that is predictive on fresh samples:

$$\|\widetilde{f} - f^*\|_D^2 := \mathop{\mathbb{E}}_{x \sim D}[|\widetilde{f}(x) - f^*(x)|^2] \le \varepsilon \cdot \mathop{\mathbb{E}}_{x,y}[|y - f^*(x)|^2] \tag{2}$$

where $f^* \in \mathcal{F}$ minimizes that RHS. The polynomial regression problem is when $\mathcal{F}$ is the set of degree-$(d-1)$ polynomials in the limit as $n \to \infty$, since we know the distribution $D$ and can query any point in it.

We state our theorem here in two cases: when $y_i$ is an unbiased estimator for $f(x_i)$ for each $x_i$, in which case $\widetilde{f}$ converges to $f = f^*$; and when $y_i$ is biased, in which case $\widetilde{f}$ converges to $f^*$ but not necessarily $f$.

**Theorem 1.2.** *Let $\mathcal{F}$ be a linear family of functions from a domain $G$ to $\mathbb{C}$ with dimension $d$, and consider any (unknown) distribution on $(x, y)$ over $G \times \mathbb{C}$. Let $D$ be the marginal distribution over $x$, and suppose it has bounded "condition number"*

$$K := \sup_{h \in \mathcal{F}: h \neq 0} \frac{\sup_{x \in G} |h(x)|^2}{\|h\|_D^2}. \tag{3}$$

*Let $f^* \in \mathcal{F}$ minimize $\mathbb{E}[|f(x) - y|^2]$. For any $\varepsilon < 1$, there exists an efficient randomized algorithm that takes $O(K \log d + \frac{K}{\varepsilon})$ unlabeled samples from $D$ and requires $O(\frac{d}{\varepsilon})$ labels to output $\widetilde{f}$ such that*

$$\mathop{\mathbb{E}}_{\widetilde{f}} \mathop{\mathbb{E}}_{x \sim D}[|\widetilde{f}(x) - f^*(x)|^2] \leq \varepsilon \cdot \mathop{\mathbb{E}}_{x,y}[|y - f^*(x)|^2].$$

A few points are in order. First, notice that if we merely want to optimize the number of labels, it is possible to take infinite number of samples from $D$ to learn it and then query whatever desired labels on $x \in \mathsf{supp}(D)$. This is identical to the query access model, where $\Theta(d/\varepsilon)$ queries is necessary and sufficient from Theorem 1.1. On the other hand, if we focus on unlabeled sample complexity, a natural solution is to query every sample point and calculating the ERM $\widetilde{f}$; one can show that this takes $\Theta(K \log d + K/\varepsilon)$ samples [CDL13]. Thus both the unlabeled and labeled sample complexity of our algorithm are optimal up to a constant factor.

Finally, in settings with a "true" signal $f(x)$ one may want $\widetilde{f} \approx f$ rather than $\widetilde{f} \approx f^*$. Such a result follows directly from the Pythagorean theorem, although (if the noise is biased, so $f^* \neq f$) the approximation becomes $(1 + \varepsilon)$ rather than $\varepsilon$:

**Corollary 1.3.** *Suppose that $y(x) = f(x) + g(x)$, where $f \in \mathcal{F}$ is the "true" signal and $g$ is arbitrary and possibly randomized "noise". Then in the setting of Theorem 1.2, with $\|\cdot\|_D$ defined as in (2),*

1. *$\mathbb{E}[\|\widetilde{f} - f\|_D^2] \leq \varepsilon \cdot \mathbb{E}[\|g\|_D^2]$, if each $g(x)$ is a random variable with $\mathbb{E}_{x,g}[g(x)] = 0$.*
2. *Otherwise, $\|\widetilde{f} - f\|_D \leq (1 + O(\varepsilon)) \cdot \|g\|_D$ with probability 0.99.*

To make the result concrete, we present the following implication:

**Example 1.4.** *Consider fitting $n$-variate degree-$d$ polynomials on $[-1, 1]^n$. There are $\binom{n+d}{d}$ monomials in the family, so Theorem 1.2 shows that querying $O(\binom{n+d}{d})$ points can achieve a constant-factor approximation to the optimal polynomial. By contrast, uniform sampling would work well for low $d$, but loses a poly($d$) factor; Chebyshev sampling would work well for low $n$, but loses a $2^{O(n)}$ factor; leverage score sampling would lose a $\log \binom{n+d}{d}$ factor.*

**Continuous Sparse Fourier transform.** Next we study sampling methods for learning a non-linear family: $k$-Fourier-sparse signals in the continuous domain. We consider the family of bandlimited $k$-Fourier-sparse signals

$$\mathcal{F} = \left\{ f(x) = \sum_{j=1}^{k} v_j \cdot e^{2\pi \boldsymbol{i} f_j x} \middle| f_j \in \mathbb{R} \cap [-F, F], C_j \in \mathbb{C} \right\} \tag{4}$$

over the domain $D$ uniform on $[-1, 1]$.

Because the frequencies $f_j$ can be any real number in $[-F, F]$, this family is not well conditioned. If all $f_j \to 0$, a Taylor approximation shows that one can arbitrarily approximate any degree $(k-1)$ polynomial; hence $K$ in (3) is at least $\Theta(k^2)$.

To improve the sample complexity of learning $\mathcal{F}$, we apply importance sampling for it by biasing $x \in [-1, 1]$ proportional to the largest variance at each point: $\sup_{f \in \mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2}$. This is a natural extension of leverage score sampling, since it matches the leverage score distribution when $\mathcal{F}$ is linear. Our main contribution is a simple upper bound that closely approximates the importance sampling weight for $k$-Fourier-sparse signals at every point $x \in (-1, 1)$.

**Theorem 1.5.** *For any $x \in (-1, 1)$,*

$$\sup_{f \in \mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2} \lesssim \frac{k \log k}{1 - |x|}.$$

Combining this with the condition number bound $K = \tilde{O}(k^4)$ in [CKPS16], this gives an explicit sampling distribution with a "reweighted" condition number (as defined in Section 2) of $O(k \log^2 k)$; this is almost tight, since $k$ is known to be necessary. We show the weight density in Figure 1.

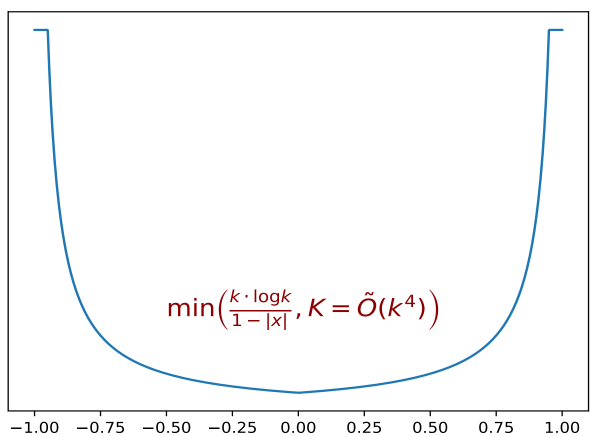


Figure 1: Explicit weights for $k$-Fourier-sparse signals

The reweighted condition number indicates that $m = \tilde{O}(k)$ suffices for the empirical estimation of $\|f\|_D$ for any fixed $f \in \mathcal{F}$. We show that this implies that $m = \tilde{O}(k^4 + k^2 \cdot \log F)$ guarantees the empirical estimation of *all* $f \in \mathcal{F}$, so that the ERM $\widetilde{f} \approx f$. (The extra loss of $m$ is due to the infinitely many possible frequencies in this family). This is a much better polynomial in $k$ than was previously known to be possible for the problem [CKPS16]. We believe that this sampling approach directly translates to improvements in the polynomial time recovery algorithm of [CKPS16], but that algorithm is quite complicated so we leave this for future work.

## 1.1 Related Work

**Linear regression.** A large body of work considers ways to subsample linear regression problems so the solution $\widetilde{\beta}$ to the subsampled problem approximates the overall solution $\beta^*$. The most common such method is leverage score sampling, which achieves the guarantee of Theorem 1.1 with $O(d \log d + d/\varepsilon)$ samples [DMM08, MI10, Mah11, Woo14].

Several approaches have attempted to go beyond this $O(d \log d)$ sample complexity. Both [BDMI13] and [SWZ19] apply the deterministic linear-sample spectral sparsification of [BSS12] to the matrix $n \times (d+1)$ matrix $[X|y]$, to find a size $O(d/\varepsilon)$ set that would suffice for Theorem 1.1. However, this procedure requires knowing the entirety of $y$ to find $S$, so it does not help for active learning. [AZLSW17] showed how such a procedure can additionally have a number of extra properties, such as that each sample has equal weight. However, all these results involve deterministic sampling procedures, so cannot tolerate adversarial noise.

Another line of research on minimizing the query complexity of linear regression is volume sampling, which samples a set of points proportional to the volume of the parallelepiped spanned by an orthonormal basis on these points. Recently, [DW17] showed that exactly $d$ points chosen from volume sampling can achieve the guarantee of Theorem 1.1, except with an approximation ratio $d+1$ rather than $1+O(\epsilon)$. In a subsequent work, [DWH18] showed that standard volume sampling would need $\Omega(K)$ samples to achieve any constant approximation ratio, but that a variant of volume sampling can match leverage score sampling with $O(d \log d + d/\varepsilon)$ samples.

| | | # queries | Remark |
|---|---|---|---|
| Uniform sampling | [CDL13, HS16] | $O(K \log d)$ | |
| Leverage score sampling | [Mah11] | $\Theta(d \log d)$ | |
| Boutsidis et al. | [BDMI13] | $\lvert \mathsf{supp}(D) \rvert$ | only needs $O(d)$ points in (2) |
| Volume sampling | [DW17] | $d$ | for $\varepsilon = d+1$ |
| | [DWH18] | $\Omega(K)$ | for $\varepsilon = 1.5$ |
| Rescaled volume sampling | [DWH18] | $O(d \log d)$ | |
| This work | | $O(d)$ | |

Table 1: Summary on the sample complexity of learning linear families, for $\varepsilon = \Theta(1)$ unless otherwise specified. Note that $K \geq d$.

For active regression, [SM14] provide an algorithm of $O(d \log d)^{5/4}$ labels to achieve the desired guarantee of ERM, while they do not give an explicit bound on the number of unlabeled points in the algorithm. [CKNS15] propose a different approach assuming additional structure for the distribution $D$ and knowledge about the noise $g$, allowing stronger results than are possible in our setting.

One linear family of particular interest is univariate polynomials of degree $d$ with the uniform distribution over $[-1,1]$. [CDL13] show that $O(K \log d)$ samples suffices for (2) for any linear family. In particular, they prove $m = O(d \log d)$ samples generated from the Chebyshev weight are sufficient, because it is the limit of the leverage scores of univarite polynomials. [CKPS16] avoids the extra loss $\log d$ by generating every point $x_i$ using a distinct distribution: it partitions the Chebyshev weight into $O(d)$ intervals of equal summations and sample one point from each interval. However, this partition may not exist for arbitrary linear families and distributions.

**Sparse Fourier transform.** There is a long line of research on sparse Fourier transform in the continuous setting, e.g., Prony's method from 1795, Hilbert's inequality by [MV74] and Matrix Pencil method [BM86, Moi15] to name a few. At the same time, less is known about the worst case guarantees without any assumption on separation between the frequencies; this depends on the condition number $K$, which is between $k^2$ and $\tilde{O}(k^4)$ as noted above [CKPS16]. We note in passing the bound on $K$ and Theorem 1.5 is analogous to Markov Brothers' inequality and the

Bernstein inequality for univariate polynomials.

A number of works have studied importance sampling for sparse recovery and sparse Fourier transforms. [RW12] considered the case where $\mathcal{F}$ is sparse in a well-behaved orthonormal basis such as polynomials sparse in Legendre basis using the Chebyshev distribution. We refer to the survey [War15] for a detailed discussion. Recently, [AKM+17] give a study about kernel ridge scores for signals with known Fourier transform structures such as the Gaussian kernel in multi-dimension. However, the weight shown in [AKM+17] is not close to optimal for multi-dimension, while our weight is almost tight.

**Organization.** We introduce our approaches and "well-balanced" procedures and outline the proofs of our results in Section 2. After introducing notation and tools in Section 3, we prove a "well-balanced" procedure guarantees (2) with high probability. Then we show the randomized spectral sparsification of [LS15] is "well-balanced" in Section 5. For completeness, we analyze the number of samples generated by one distribution in Section 6. Next we combine the results of the previous two sections to prove our results about active learning in Section 7. We show information lower bound on the sample complexity in Section 8. Finally, we prove our results about sparse Fourier transform in Section 9.

## 2 Proof Overview

We present our proof sketch in the notation of Theorem 1.2. Consider observations of the form $y(x) = f(x) + g(x)$ for $f$ in a (not yet necessarily linear) family $\mathcal{F}$ and $g$ an arbitrary, possibly random function.

**Improved conditioning by better sampling.** We start with the noiseless case of $g = 0$ in the query access model, and consider the problem of estimating $\|y\|_D^2 = \|f\|_D^2$ with high probability. If we sample points $x_i \sim D'$ for some distribution $D'$, then we can estimate $\|f\|_D^2$ as the empirical norm

$$\frac{1}{m}\sum_{i=1}^{m} \frac{D(x_i)}{D'(x_i)} |f(x_i)|^2 \tag{5}$$

which has the correct expectation. To show the expectation concentrates, we should bound the maximum value of the summand, which we define to be the "reweighted" condition number

$$K_{D'} = \sup_{x} \Big\{ \sup_{f\in\mathcal{F}} \{ \frac{D(x)}{D'(x)} \cdot \frac{|f(x)|^2}{\|f\|_D^2} \} \Big\}.$$

We define $D_{\mathcal{F}}$ to minimize this quantity, by making the inner term the same for every $x$. Namely, we pick

$$D_{\mathcal{F}}(x) = \frac{1}{\kappa} D(x) \cdot \sup_{f\in\mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2} \qquad \text{for } \kappa = \mathop{\mathbb{E}}_{x\sim D}\left[ \sup_{f\in\mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2} \right]. \tag{6}$$

This shows that by sampling from $D_{\mathcal{F}}$ rather than $D$, the condition number of our estimate (5) improves from $K = \sup_{x\in \mathsf{supp}(D)} \big\{ \sup_{f\in\mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2} \big\}$ to $\kappa = \mathop{\mathbb{E}}_{x\sim D}\left[ \sup_{f\in\mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2} \right]$.

From the Chernoff bound, $O(\frac{\kappa \cdot \log \frac{1}{\delta}}{\varepsilon^2})$ samples from $D_{\mathcal{F}}$ let us estimate $\|f\|_D^2$ to within accuracy $1 \pm \varepsilon$ with probability $1-\delta$ for any fixed function $f \in \mathcal{F}$. To be able to estimate *every* $f \in \mathcal{F}$, a basic solution would be to apply a union bound over an $\varepsilon$-net of $\mathcal{F}$.

Linear function families $\mathcal{F}$ let us improve the result in two ways. First, we observe that $\kappa = d$ for any dimension $d$ linear function space; in fact, $D_{\mathcal{F}}$ is the leverage score sampling distribution. Second, we can replace the union bound by a matrix concentration bound, showing that $O(\frac{d \log \frac{d}{\delta}}{\varepsilon^2})$ samples from $D_{\mathcal{F}}$ suffice to estimate $\|f\|_D^2$ to within $1 \pm \varepsilon$ for all $f \in \mathcal{F}$ with probability $1-\delta$. However, this approach needs $\Omega(d \log d)$ samples due to a coupon-collector argument, because it only samples points from one distribution $D_{\mathcal{F}}$.

**The effect of noise.** The spectral sparsifier given by [BSS12] could replace the matrix concentration bound above, and estimate $\|f\|_D^2$ for every $f \in \mathcal{F}$ with only $O(d)$ samples. The issue with this is that it would not be robust against adversarial noise, because the sample points $x_i$ are deterministic. Now we consider our actual problem, which is to estimate $f$ from $y = f + g$ for nonzero noise $g$ of $\mathbb{E}[\|g\|_D^2] = \sigma^2$. We need our sample points to both be sampled non-independently (to avoid coupon-collector issues) but still fairly randomly (so adversarial noise cannot predict it). A natural approach is to design a sequence of distributions $D_1, \cdots, D_m$ ($m$ is not necessarily fixed) then sample $x_i \sim D_i$ and assign a weight $w_i$ for $x_i$, where $D_{i+1}$ could depend on the previous points $x_1, \cdots, x_i$.

Given samples $(x_1, \cdots, x_m)$ with weights $(w_1, \cdots, w_m)$, the empirical risk minimizer is the function $\widetilde{f} \in \mathcal{F}$ closest to $y$ under the empirical norm $\sum\limits_{i \in [m]} w_i \cdot |f(x_i)|^2$. When $\mathcal{F}$ is a linear family, the solution $\widetilde{f}$ is a linear projection, so it acts on $f$ and $g$ independently. If the empirical norm is a good estimator for $\mathcal{F}$, the projection of $f \in \mathcal{F}$ into the linear subspace $\mathcal{F}$ equals $f$. Hence the error $\widetilde{f} - f$ is the projection of $g$ onto $\mathcal{F}$ under the empirical norm.

First, suppose that $g$ is orthogonal to $\mathcal{F}$ under the true norm $\|\cdot\|_D$—for instance, if $g(x)$ is an independent mean-zero random variable for each $x$. In this case, the expected value of the projection of $g$ is zero. At the same time, we can bound the variance of the projection of a single random sample of $g$ drawn from $D_i$ by the condition number $K_{D_i} \cdot \sigma^2$. Ideally each $K_{D_i}$ would be $O(d)$, but we do not know how to produce such distributions while still getting linear sample spectral sparsification. Therefore we use a coefficient $\alpha_i$ to control every $K_{D_i}$, and set $w_i = \alpha_i \cdot \frac{D(x_i)}{D_i(x_i)}$ instead of $\frac{D(x_i)}{m D_i(x_i)}$. The result is that—if $\sum_i \alpha_i = O(1)$—the projection of the noise has variance $O\big(\max\limits_{i \in [m]}\{\alpha_i K_{D_i}\}\big) \cdot \sigma^2$. This motivates our definition of "well-balanced" sampling procedures:

**Definition 2.1.** *Given a linear family $\mathcal{F}$ and underlying distribution $D$, let $P$ be a random sampling procedure that terminates in $m$ iterations ($m$ is not necessarily fixed) and provides a coefficient $\alpha_i$ and a distribution $D_i$ to sample $x_i \sim D_i$ in every iteration $i \in [m]$.*

*We say $P$ is an* $\varepsilon$-well-balanced sampling procedure *if it satisfies the following two properties:*

1. *With probability* 0.9*, for weight $w_i = \alpha_i \cdot \frac{D(x_i)}{D_i(x_i)}$ of each $i \in [m]$,*
$$\sum_{i=1}^{m} w_i \cdot |h(x_i)|^2 \in \left[\frac{3}{4}, \frac{5}{4}\right] \cdot \|h\|_D^2 \quad \forall h \in \mathcal{F}.$$
*Equivalently (as shown in Lemma 4.2 in Section 4.1), given any orthonormal basis $v_1, \ldots, v_d$ of $\mathcal{F}$ under $D$, the matrix $A(i,j) = \sqrt{w_i} \cdot v_j(x_i) \in \mathbb{C}^{m \times d}$ has $\lambda(A^* A) \in [\frac{3}{4}, \frac{5}{4}]$.*

2. *The coefficients always have $\sum_i \alpha_i \leq \frac{5}{4}$ and $\alpha_i \cdot K_{D_i} \leq \epsilon/2$.*

Intuitively, the first property says that the sampling procedure preserves the signal, and the second property says that the recovery algorithm does not blow up the noise on average. For such a sampling procedure we consider the ERM from its execution as follows.

Our definition of a well-balanced sampling procedure allows property 1 to fail 10% of the time, but our algorithm will only perform well in expectation when property 1 is satisfied. Therefore we rerun the sampling procedure until it has a "good" execution that satisfies property 1.

**Definition 2.2.** *Given a* well-balanced sampling procedure $P$*, we say one execution of* $P$ *is* good *only if the samples* $x_i$ *with weights* $w_i = \alpha_i \cdot \frac{D(x_i)}{D_i(x_i)}$ *satisfy the first property in Definition 2.1, which can be checked efficiently by calculating* $\lambda(A^*A)$.

*Given a joint distribution* $(D, Y)$ *and an execution of a* well-balanced sampling procedure $P$ *with* $x_i \sim D_i$ *and* $w_i = \alpha_i \cdot \frac{D(x_i)}{D_i(x_i)}$ *of each* $i \in [m]$*, let the weighted ERM of this execution be* $\widetilde{f} = \arg\min_{h \in \mathcal{F}} \left\{ \sum_{i=1}^{m} w_i \cdot |h(x_i) - y_i|^2 \right\}$ *by querying* $y_i \sim (Y|x_i)$ *for each point* $x_i$ *.*

In Section 4 we prove that $\widetilde{f}$ satisfies the desired guarantee, which implies Theorem 1.1.

**Theorem 2.3.** *Given a linear family* $\mathcal{F}$*, joint distribution* $(D, Y)$*, and* $\varepsilon > 0$*, let* $P$ *be an* $\varepsilon$*-well-balanced sampling procedure for* $\mathcal{F}$ *and* $D$*, and let* $f = \arg\min_{h \in \mathcal{F}} \mathbb{E}_{(x,y)\sim(D,Y)}[|y - h(x)|^2]$ *be the true risk minimizer. Then the weighted ERM* $\widetilde{f}$ *of a good execution of* $P$ *satisfies*

$$\|f - \widetilde{f}\|_D^2 \leq \epsilon \cdot \mathbb{E}_{(x,y)\sim(D,Y)}[|y - f(x)|^2] \text{ in expectation.}$$

For a noise function $g$ not orthogonal to $\mathcal{F}$ in expectation, let $g^{\perp}$ and $g^{\parallel}$ denote the decomposition of $g$ where $g^{\perp}$ is the orthogonal part and $g^{\parallel} = g - g^{\perp} \in \mathcal{F}$. The above theorem indicates $\|\widetilde{f} - f\|_D \leq \|g^{\parallel}\|_D + \sqrt{\varepsilon} \cdot \|g^{\perp}\|_D$, which gives $(1+\varepsilon)\|g\|_D$-closeness via the Pythagorean theorem. This result appears in Corollary 4.1 of Section 4.

**Well-balanced sampling procedures.** We observe that two standard sampling procedures are well-balanced, so they yield agnostic recovery guarantees by Theorem 2.3. The simplest approach is to set each $D_i$ to be a fixed distribution $D'$ and $\alpha_i = 1/m$ for all $i$. For $m = O(K_{D'} \log d + K_{D'}/\varepsilon)$, this gives an $\varepsilon$-well-balanced sampling procedure. These results appear in Section 6.

We get a stronger result of $m = O(d/\varepsilon)$ using the randomized BSS algorithm from [LS15]. The [LS15] algorithm iteratively chooses points $x_i$ from distributions $D_i$. A term considered in their analysis—the largest increment of eigenvalues—is equivalent to our $K_{D_i}$. By looking at the potential functions in their proof, we can extract coefficients $\alpha_i$ bounding $\alpha_i K_{D_i}$ in our setting. This lets us show that the algorithm is a well-balanced sampling procedure; we do so in Section 5.

**Active learning.** Next we consider the active learning setting, where we don't know the distribution $D$ and only receive samples $x_i \sim D$, but can choose which $x_i$ receive labels $y_i$. Let $K$ be the condition number of the linear family $\mathcal{F}$. Our algorithms uses $n = O(K \log d + \frac{K}{\varepsilon})$ unlabeled samples and $m = O(\frac{d}{\varepsilon})$ labeled samples, and achieves the same guarantees as in the query access model.

For simplicity, we start with $g$ orthogonal to $\mathcal{F}$ under $\|\cdot\|_D$. At first, let us focus on the number of *unlabeled* points. We could take $n = O(K \log d + \frac{K}{\varepsilon})$ points from $D$ and request the label of each point $x_i$. By Theorem 2.3 with the simpler well-balanced sampling procedure mentioned above using $D' = D$, the ERM $f'$ on these $n$ points is $\epsilon \cdot \mathbb{E}_D[|g(x)|^2]$-close to $f$.

Then let us optimize the number of labeled samples. For $n$ random points from $D$, let $D_0$ denote the uniform measurement on these points. Although we cannot apply the linear-sample well-balanced sampling procedure $P$ to the unknown $D$, we can apply it to $D_0$. By Theorem 2.3, the ERM $\widetilde{f}$ of $P$ on $D_0$ satisfies $\|\widetilde{f}-f'\|_{D_0}^2 \le \epsilon \cdot \mathop{\mathbb{E}}_{x\sim D_0}[|y(x)-f'(x)|^2]$. By the triangle inequality and the fact that $D_0$ is an good empirical estimation of $\mathcal{F}$ under measurement $D$, this gives $\|f-\widetilde{f}\|_D^2 \lesssim \epsilon \cdot \mathop{\mathbb{E}}_{D}[|g(x)|^2]$.

Notice that $f'$ only appears in the analysis and we do not need it in the calculation of $\widetilde{f}$ given $D_0$. By rescaling a constant factor of $\varepsilon$, this leads to the following theorem proved in Section 7.

**Theorem 2.4.** *Consider any dimension $d$ linear space $\mathcal{F}$ of functions from a domain $G$ to $\mathbb{C}$. Let $(D,Y)$ be a joint distribution over $G\times\mathbb{C}$ and $f=\arg\min_{h\in\mathcal{F}} \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y-h(x)|^2]$.*

*Let $K=\sup_{h\in\mathcal{F}:h\neq 0}\frac{\sup_{x\in G}|h(x)|^2}{\|h\|_D^2}$. For any $\varepsilon>0$, there exists an efficient algorithm that takes $O(K\log d+\frac{K}{\varepsilon})$ unlabeled samples from $D$ and requests $O(\frac{d}{\varepsilon})$ labels to output $\widetilde{f}$ satisfying*

$$\mathop{\mathbb{E}}_{x\sim D}[|\tilde{f}(x)-f(x)|^2] \le \varepsilon \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y-f(x)|^2] \textit{ in expectation.}$$

**Lower bounds.** We first prove a lower bound on the query complexity using information theory. The Shannon-Hartley Theorem indicates that under the i.i.d. Gaussian noise $N(0,1/\varepsilon)$, for a function $f$ with $|f(x)|\le 1$ at every point $x$, any observation $y(x)=f(x)+N(0,1/\varepsilon)$ obtains $O(\varepsilon)$ information about $f$. Because the dimension of $\mathcal{F}$ is $d$, this indicates $\Omega(d/\varepsilon)$ queries is necessary to recover a function in $\mathcal{F}$.

Next, for any $K$, $d$, and $\varepsilon$ we construct a distribution $D$ and dimension-$d$ linear family $\mathcal{F}$ with condition number $K$ over $D$, such that the sample complexity of achieving (2) is $\Omega(K\log d+K/\varepsilon)$. The first term comes from the coupon collector problem, and the second comes from the above query bound. We summarize the upper bounds and lower bounds for sample complexity and query complexity in Table 2.

| | Optimal value | Lower bound | Upper bound |
|---|---|---|---|
| Query complexity | $\Theta(d/\varepsilon)$ | Theorem 8.1 | Theorem 1.1 |
| Sample complexity | $\Theta(K\log d+K/\varepsilon)$ | Theorem 8.4 | Theorem 6.3 |

Table 2: Lower bounds and upper bounds in different access models

**Signals with $k$-sparse Fourier transform.** We now consider the nonlinear family $\mathcal{F}$ of functions with $k$-sparse Fourier transform defined in (4), over the distribution $D=[-1,1]$. As discussed at (6), even for nonlinear function families, sampling from $D_{\mathcal{F}}$ proportional to $\sup_{f\in\mathcal{F}}\frac{|f(x)|^2}{\|f\|_D^2}$ improves the condition number from $K$ to $\kappa=\mathop{\mathbb{E}}_{x\in D}\sup_{f\in\mathcal{F}}\frac{|f(x)|^2}{\|f\|_D^2}$, which is $\tilde{O}(k)$ given Theorem 1.5 and $K=\tilde{O}(k^4)$.

Before sketching the proof of Theorem 1.5, let us revisit the $\widetilde{O}(k^4)$ bound for $K$ shown in [CKPS16]. The key step—Claim 5.2 in [CKPS16]—showed that for any $\Delta>0$ and $f\in\mathcal{F}$, $f(x)$ can be expressed as a linear combination of $\{f(x+j\Delta)\mid j=1,\ldots,l\}$ with constant coefficients and $l=\tilde{O}(k^2)$. This upper bounds $|f(-1)|^2$ in terms of $|f(-1+\Delta)|^2+\cdots+|f(-1+l\cdot\Delta)|^2$ and then $|f(-1)|^2/\|f\|_D^2$ by integrating $\Delta$ from 0 to $2/l$.

The improvement of Theorem 1.5 contains two steps. In the first step, we show that $f(x)$ can be expressed as a constant-coefficient linear combination of the elements of an $O(k)$-length arithmetic sequence on both sides of $x$, namely, $\{f(x-2k\cdot\Delta),\ldots,f(x+2k\cdot\Delta)\}\setminus f(x)$. This is much shorter than the $\widetilde{O}(k^2)$ elements required by [CKPS16] for the one-sided version, and provides an $\widetilde{O}(k^2)$

factor improvement. Next we find $k$ such linear combinations that are almost orthogonal to each other to remove the extra $k$ factor. These two let us show that

$$\sup_{f\in\mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2} = O\left(\frac{k\log k}{1-|x|}\right)$$

for any $x \in (-1,1)$. This leads to $\kappa = O(k\log^2 k)$, which appears in Theorem 9.1 of Section 9.

# 3 Notation

We use $[k]$ to denote the subset $\{1,2,\ldots,k\}$ and $1_E \in \{0,1\}$ to denote the indicator function of an event $E$.

For a vector $\vec{v} = \big(v(1),\ldots,v(m)\big) \in \mathbb{C}^m$, let $\|\vec{v}\|_k$ denote the $\ell_k$ norm, i.e., $\big(\sum_{i\in[m]} |v(i)|^k\big)^{1/k}$.

Given a self-adjoint matrix $A \in \mathbb{C}^{m\times m}$, let $\|A\|$ denote the operator norm $\|A\| = \max_{\vec{v}\neq\vec{0}} \frac{\|A\vec{v}\|_2}{\|\vec{v}\|_2}$ and $\lambda(A)$ denote all eigenvalues of $A$. For convenience, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest eigenvalue and the largest eigenvalue of $A$. Given a matrix $B$, let $B^*$ denote the conjugate transpose of $B$, i.e., $B^*(j,i) = \overline{B(i,j)}$.

Given a function $f$ with domain $G$ and a distribution $D$ over $G$, we use $\|f\|_D$ to denote the expected $\ell_2$ norm of $f(x)$ where $x \sim D$, i.e., $\|f\|_D = \big(\mathop{\mathbb{E}}_{x\sim D}[|f(x)|^2]\big)^{1/2}$. Given a sequence $S = (x_1,\ldots,x_m)$ (allowing repetition in $S$) and corresponding weights $(w_1,\ldots,w_m)$, let $\|f\|_{S,w}^2$ denote the weighted $\ell_2$ norm $\sum_{j=1}^m w_j \cdot |f(x_j)|^2$. For convenience, we omit $w$ if it is a uniform distribution on $S$, i.e., $\|f\|_S = \big(\mathbb{E}_{i\in[m]}\left[|f(x_i)|^2\right]\big)^{1/2}$.

**Weights between different distributions.** Given a distribution $D$, to estimate $\|h\|_D^2$ of a function $h$ through random samples from $D'$, we use the following notation to denote the re-weighting of $h$ between $D'$ and $D$.

**Definition 3.1.** *For any distribution $D'$ over the domain $G$ and any function $h : G \to \mathbb{C}$, let $h^{(D')}(x) = \sqrt{\frac{D(x)}{D'(x)}} \cdot h(x)$ such that $\mathop{\mathbb{E}}_{x\sim D'}\left[|h^{(D')}(x)|^2\right] = \mathop{\mathbb{E}}_{x\sim D'}\left[\frac{D(x)}{D'(x)}|h(x)|^2\right] = \mathop{\mathbb{E}}_{x\sim D}\left[|h(x)|^2\right]$. When the family $\mathcal{F}$ and D is clear, we use $K_{D'}$ to denote the condition number of sampling from $D'$, i.e.,*

$$K_{D'} = \sup_x \left\{ \sup_{h\in\mathcal{F}} \left\{ \frac{|h^{(D')}(x)|^2}{\|h^{D'}\|_{D'}^2} \right\} \right\} = \sup_x \left\{ \frac{D(x)}{D'(x)} \cdot \sup_{h\in\mathcal{F}} \{ \frac{|h(x)|^2}{\|h\|_2^2} \} \right\}.$$

By the same reason, for a random sample $x$ from distribution $D'$, we always use $w_x = \frac{D(x)}{D'(x)}$ to re-weight the sample $x$ such that it keeps the same expectation:

$$\mathop{\mathbb{E}}_{x\sim D'}[w_x \cdot |h(x)|^2] = \mathop{\mathbb{E}}_{x\sim D'}[\frac{D(x)}{D'(x)} \cdot |h(x)|^2] = \mathop{\mathbb{E}}_{x\sim D}[|h(x)|^2] = \|h\|_D^2.$$

# 4 Recovery Guarantee for Well-Balanced Samples

In this section, we show for well-balanced sampling procedures (per Definition 2.1) that the weighted ERM of a good execution (per Definition 2.2) approximates the true risk minimizer, and hence the true signal. For generality, we first consider points and labels from a joint distribution $(D,Y)$.

**Theorem 2.3.** *Given a linear family $\mathcal{F}$, joint distribution $(D, Y)$, and $\varepsilon > 0$, let $P$ be an $\varepsilon$-well-balanced sampling procedure for $\mathcal{F}$ and $D$, and let $f = \arg\min_{h \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - h(x)|^2]$ be the true risk minimizer. Then the weighted ERM $\widetilde{f}$ of a good execution of $P$ satisfies*

$$\|f - \widetilde{f}\|_D^2 \leq \epsilon \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2] \text{ in expectation.}$$

Next, we provide a corollary for specific kinds of noise. In the first case, we consider noise functions representing independently mean-zero noise at each position $x$ such as i.i.d. Gaussian noise. Second, we consider arbitrary noise functions on the domain.

**Corollary 4.1.** *Given a linear family $\mathcal{F}$ and distribution $D$, let $y(x) = f(x) + g(x)$ for $f \in \mathcal{F}$ and $g$ a randomized function. Let $P$ be an $\varepsilon$-well-balanced sampling procedure for $\mathcal{F}$ and $D$. The weighted ERM $\widetilde{f}$ of a good execution of $P$ satisfies*

1. *$\|\tilde{f} - f\|_D^2 \leq \varepsilon \cdot \mathop{\mathbb{E}}_g[\|g\|_D^2]$ in expectation, when $g(x)$ is a random function from $G$ to $\mathbb{C}$ where each $g(x)$ is an independent random variable with $\mathop{\mathbb{E}}_g[g(x)] = 0$.*

2. *With probability 0.99, $\|\tilde{f} - f\|_D \leq (1 + \varepsilon) \cdot \|g\|_D$ for any other noise function $g$.*

In the rest of this section, we prove Theorem 2.3 in Section 4.1 and Corollary 4.1 in Section 4.2. We discuss the speedup of the calculation of the ERM and show the fast polynomial regression in Section 4.3.

## 4.1 Proof of Theorem 2.3

We introduce a few more notation in this proof. Given $\mathcal{F}$ and the measurement $D$, let $\{v_1, \ldots, v_d\}$ be a fixed orthonormal basis of $\mathcal{F}$, where inner products are taken under the distribution $D$, i.e., $\mathop{\mathbb{E}}_{x\sim D}[v_i(x) \cdot \overline{v_j(x)}] = 1_{i=j}$ for any $i, j \in [d]$. For any function $h \in \mathcal{F}$, let $\alpha(h)$ denote the coefficients $(\alpha(h)_1, \ldots, \alpha(h)_d)$ under the basis $(v_1, \ldots, v_d)$ such that $h = \sum_{i=1}^d \alpha(h)_i \cdot v_i$ and $\|\alpha(h)\|_2 = \|h\|_D$.

We characterize the first property in Definition 2.1 of *well-balanced sampling procedures* as bounding the eigenvalues of $A^* \cdot A$, where $A$ is the $m \times d$ matrix defined as $A(i, j) = \sqrt{w_i} \cdot v_j(x_i)$.

**Lemma 4.2.** *For any $\varepsilon > 0$, given $S = (x_1, \ldots, x_m)$ and their weights $(w_1, \ldots, w_m)$, let $A$ be the $m \times d$ matrix defined as $A(i, j) = \sqrt{w_i} \cdot v_j(x_i)$. Then*

$$\|h\|_{S,w}^2 := \sum_{j=1}^m w_j \cdot |f(x_j)|^2 \in [1 \pm \varepsilon] \cdot \|h\|_D^2 \quad \textit{for every } h \in \mathcal{F}$$

*if and only if the eigenvalues of $A^*A$ are in $[1 - \varepsilon, 1 + \varepsilon]$.*

*Proof.* Notice that

$$A \cdot \alpha(h) = \big(\sqrt{w_1} \cdot h(x_1), \ldots, \sqrt{w_m} \cdot h(x_m)\big). \tag{7}$$

Because

$$\|h\|_{S,w}^2 = \sum_{i=1}^m w_i |h(x_i)|^2 = \|A \cdot \alpha(h)\|_2^2 = \alpha(h)^* \cdot (A^* \cdot A) \cdot \alpha(h) \in [\lambda_{\min}(A^* \cdot A), \lambda_{\max}(A^* \cdot A)] \cdot \|h\|_D^2$$

and $h$ is over the linear family $\mathcal{F}$, these two properties are equivalent.

□

Next we consider the calculation of the weighted ERM $\widetilde{f}$. Given the weights $(w_1, \cdots, w_m)$ on $(x_1, \ldots, x_m)$ and labels $(y_1, \ldots, y_m)$, let $\vec{y}_w$ denote the vector of weighted labels $(\sqrt{w_1} \cdot y_1, \ldots, \sqrt{w_m} \cdot y_m)$. From (7), the empirical distance $\|h - (y_1, \ldots, y_m)\|_{S,w}^2$ equals $\|A \cdot \alpha(h) - \vec{y}_w\|_2^2$ for any $h \in \mathcal{F}$. The function $\widetilde{f}$ minimizing $\|h - (y_1, \ldots, y_m)\|_{S,w} = \|A \cdot \alpha(h) - \vec{y}_w\|_2$ overall all $h \in \mathcal{F}$ is the pseudoinverse of $A$ on $\vec{y}_w$, i.e.,

$$\alpha(\widetilde{f}) = (A^* \cdot A)^{-1} \cdot A^* \cdot \vec{y}_w \text{ and } \widetilde{f} = \sum_{i=1}^{d} \alpha(\widetilde{f})_i \cdot v_i.$$

Finally, we consider the distance between $f = \underset{h \in \mathcal{F}}{\arg\min}\big\{ \underset{(x,y)\sim(D,Y)}{\mathbb{E}}[|h(x) - y|^2]\big\}$ and $\widetilde{f}$. For convenience, let $\vec{f}_w = \big(\sqrt{w_1} \cdot f(x_1), \ldots, \sqrt{w_m} \cdot f(w_m)\big)$. Because $f \in \mathcal{F}$, $(A^* \cdot A)^{-1} \cdot A^* \cdot \vec{f}_w = \alpha(f)$. This implies

$$\|\widetilde{f} - f\|_D^2 = \|\alpha(\widetilde{f}) - \alpha(f)\|_2^2 = \|(A^* \cdot A)^{-1} \cdot A^* \cdot (\vec{y}_w - \vec{f}_w)\|_2^2.$$

We assume $\lambda\big((A^* \cdot A)^{-1}\big)$ is bounded and consider $\|A^* \cdot (\vec{y}_w - \vec{f}_w)\|_2^2$.

**Lemma 4.3.** *Let $P$ be an random sampling procedure terminating in $m$ iterations ($m$ is not necessarily fixed) that in every iteration $i$, it provides a coefficient $\alpha_i$ and a distribution $D_i$ to sample $x_i \sim D_i$. Let the weight $w_i = \alpha_i \cdot \frac{D(x_i)}{D_i(x_i)}$ and $A \in \mathbb{C}^{m \times d}$ denote the matrix $A(i,j) = \sqrt{w_i} \cdot v_j(x_i)$. Then for $f = \underset{h \in \mathcal{F}}{\arg\min} \underset{(x,y)\sim(D,Y)}{\mathbb{E}}[|y - h(x)|^2]$,*

$$\mathop{\mathbb{E}}_{P}\left[\|A^*(\vec{y}_w - \vec{f}_{S,w})\|_2^2\right] \le \sup_P \big\{\sum_{i=1}^{m} \alpha_i\big\} \cdot \max_j \big\{\alpha_j \cdot K_{D_j}\big\} \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - f(x)|^2],$$

*where $K_{D_i}$ is the condition number for samples from $D_i$: $K_{D_i} = \sup_x \left\{ \frac{D(x)}{D_i(x)} \cdot \sup_{v \in \mathcal{F}} \big\{ \frac{|v(x)|^2}{\|v\|_2^2} \big\} \right\}$.*

*Proof.* For convenience, let $g_j$ denote $y_j - f(x_j)$ and $\vec{g}_w \in \mathbb{C}^m$ denote the vector $\left(\sqrt{w_j} \cdot g_j|_{j=1,\ldots,m}\right) = \vec{y}_w - \vec{f}_{S,w}$ for $j \in [m]$ such that $A^* \cdot (\vec{y}_w - \vec{f}_{S,w}) = A^* \cdot \vec{g}_w$.

$$\begin{aligned}
\mathbb{E}[\|A^* \cdot \vec{g}_w\|_2^2] &= \mathbb{E}\left[\sum_{i=1}^{d} \big(\sum_{j=1}^{m} A^*(i,j)\vec{g}_w(j)\big)^2\right] \\
&= \sum_{i=1}^{d} \mathbb{E}\left[\big(\sum_{j=1}^{m} w_j \overline{v_i(x_j)} \cdot g_j\big)^2\right] = \sum_{i=1}^{d} \mathbb{E}\left[\sum_{j=1}^{m} w_j^2 \cdot |v_i(x_j)|^2 \cdot |g_j|^2\right],
\end{aligned}$$

where the last step uses the following fact

$$\mathop{\mathbb{E}}_{w_j \sim D_j}[w_j \overline{v_i(x_j)} \cdot g_j] = \mathop{\mathbb{E}}_{w_j \sim D_j}\big[\alpha_j \cdot \frac{D(x_j)}{D_j(x_j)} \overline{v_i(x_j)} g_j\big] = \alpha_j \cdot \mathop{\mathbb{E}}_{x_j \sim D, y_j \sim Y(x_j)}\big[\overline{v_i(x_j)}(y_j - f(x_j))\big] = 0.$$

We swap $i$ and $j$:

$$\begin{aligned}
\sum_{i=1}^{d} \mathbb{E}\left[\sum_{j=1}^{m} w_j^2 \cdot |v_i(x_j)|^2 \cdot |g_j|^2\right] &= \sum_{j=1}^{m} \mathbb{E}\left[\sum_{i=1}^{d} w_j |v_i(x_j)|^2 \cdot w_j |g_j|^2\right] \\
&\le \sum_{j=1}^{m} \sup_{x_j} \left\{ w_j \sum_{i=1}^{d} |v_i(x_j)|^2 \right\} \cdot \mathbb{E}\left[w_j \cdot |g_j|^2\right].
\end{aligned}$$

For $\mathbb{E}\left[w_j \cdot |g_j|^2\right]$, it equals $\underset{x_j \sim D_j, y_j \sim Y(x_j)}{\mathbb{E}}\left[\alpha_j \cdot \frac{D(x_j)}{D_j(x_j)}\left|y_j - f(x_j)\right|^2\right] = \alpha_j \cdot \underset{x_j \sim D, y_j \sim Y(x_j)}{\mathbb{E}}\left[\left|y_j - f(x_j)\right|^2\right]$.

For $\sup_{x_j}\left\{w_j \sum_{i=1}^d |v_i(x_j)|^2\right\}$, we bound it as

$$\sup_{x_j}\left\{w_j \sum_{i=1}^d |v_i(x_j)|^2\right\} = \sup_{x_j}\left\{\alpha_j \cdot \frac{D(x_j)}{D_j(x_j)} \sum_{i=1}^d |v_i(x_j)|^2\right\} = \alpha_j \sup_{x_j}\left\{\frac{D(x_j)}{D_j(x_j)} \cdot \sup_{h \in \mathcal{F}}\left\{\frac{|h(x_j)|^2}{\|h\|_D^2}\right\}\right\} = \alpha_j \cdot K_{D_j}.$$

We use the fact $\sup_{h \in \mathcal{F}}\left\{\frac{|h(x_j)|^2}{\|h\|_D^2}\right\} = \sup_{(a_1,\ldots,a_d)}\left\{\frac{|\sum_{i=1}^d a_i v_i(x_j)|^2}{\sum_{i=1}^d |a_i|^2}\right\} = \frac{(\sum_{i=1}^d |a_i|^2)(\sum_{i=1}^d |v_i(x_j)|^2)}{\sum_{i=1}^d |a_i|^2}$ by the Cauchy-Schwartz inequality. From all discussion above, we have

$$\mathbb{E}[\|A^* \cdot \vec{g}_w\|_2^2] \leq \sum_j \left(\alpha_j K_{D_j} \cdot \alpha_j \cdot \underset{(x,y)\sim(D,Y)}{\mathbb{E}}[|y - f(x)|^2]\right) \leq (\sum_j \alpha_j) \max_j \left\{\alpha_j K_{D_j}\right\} \cdot \underset{(x,y)\sim(D,Y)}{\mathbb{E}}[|y - f(x)|^2].$$

□

We combine all discussion above to prove Theorem 2.3.

*Proof of Theorem 2.3.* We assume the first property $\lambda(A^* \cdot A) \in [1 - 1/4, 1 + 1/4]$ from Definition 2.2. On the other hand, $\mathbb{E}[\|A^* \cdot (\vec{y}_w - \vec{f}_w)\|_2^2] \leq \epsilon/2 \cdot \underset{(x,y)\sim(D,Y)}{\mathbb{E}}[|y - f(x)|^2]$ from Lemma 4.3. Conditioned on the first property, we know it is still at most $\frac{\varepsilon}{2 \cdot 0.9} \cdot \underset{(x,y)\sim(D,Y)}{\mathbb{E}}[|y - f(x)|^2]$. This implies $\mathbb{E}\left[\|(A^* \cdot A)^{-1} \cdot A^* \cdot (\vec{y}_w - \vec{f}_w)\|_2^2\right] \leq \epsilon \cdot \underset{(x,y)\sim(D,Y)}{\mathbb{E}}[|y - f(x)|^2]$. □

## 4.2 Proof of Corollary 4.1

For the first part, let $(D, Y) = \big(D, f(x) + g(x)\big)$ be our joint distribution of $(x, y)$. Because the expectation $\mathbb{E}[g(x)] = 0$ for every $x \in G$, $\underset{v \in V}{\arg\min} \underset{(x,y)\sim(D,Y)}{\mathbb{E}}[|y - v(x)|^2] = f$. From Theorem 2.3, for $\alpha(\widetilde{f}) = (A^* \cdot A)^{-1} \cdot A^* \cdot \vec{y}_w$ and $m = O(d/\varepsilon)$,

$$\|\widetilde{f} - f\|_D^2 = \|\alpha(\widetilde{f}) - \alpha(f)\|_2^2 \leq \varepsilon \cdot \underset{(x,y)\sim(D,Y)}{\mathbb{E}}[|y - f(x)|^2] = \varepsilon \cdot \mathbb{E}[\|g\|_D^2], \text{ with probability } 0.99.$$

For the second part, let $g^{\|}$ be the projection of $g(x)$ to $\mathcal{F}$ and $g^{\perp} = g - g^{\|}$ be the orthogonal part to $\mathcal{F}$. Let $\alpha(g^{\|})$ denote the coefficients of $g^{\|}$ in the fixed orthonormal basis $(v_1, \ldots, v_d)$ so that $\|\alpha(g^{\|})\|_2 = \|g^{\|}\|_D$. We decompose $\vec{y}_w = \vec{f}_w + \vec{g}_w = \vec{f}_w + \vec{g^{\|}}_w + \vec{g^{\perp}}_w$. Therefore

$$\alpha(\widetilde{f}) = (A^*A)^{-1} \cdot A^* \cdot (\vec{f}_w + \vec{g^{\|}}_w + \vec{g^{\perp}}_w) = \alpha(f) + \alpha(g^{\|}) + (AA^*)^{-1}A^* \cdot \vec{g^{\perp}}_w.$$

The distance $\|\widetilde{f} - f\|_D = \|\alpha(\widetilde{f}) - \alpha(f)\|_2$ equals

$$\|(A^*A)^{-1} \cdot A^* \cdot \vec{y}_w - \alpha(f)\|_2 = \|\alpha(f) + \alpha(g^{\|}) + (A^*A)^{-1} \cdot A^* \cdot \vec{g^{\perp}}_w - \alpha(f)\|_2 = \|\alpha(g^{\|}) + (A^*A)^{-1} \cdot A^* \cdot \vec{g^{\perp}}_w\|_2.$$

From Theorem 2.3, with probability 0.99, $\|(A^*A)^{-1} \cdot A^* \cdot \vec{g^{\perp}}_w\|_2 \leq \sqrt{\varepsilon} \cdot \|g^{\perp}\|_D$. Thus

$$\begin{aligned}\|(A^*A)^{-1} \cdot A^* \cdot \vec{y}_w - \alpha(f)\|_2 &= \|\alpha(g^{\|}) + (A^*A)^{-1} \cdot A^* \cdot \vec{g^{\perp}}_w\|_2 \\ &\leq \|g^{\|}\|_D + \sqrt{\varepsilon} \cdot \|g^{\perp}\|_D.\end{aligned}$$

Let $1-\beta$ denote $\|g^{\parallel}\|_D/\|g\|_D$ such that $\|g^{\perp}\|_D/\|g\|_D = \sqrt{2\beta-\beta^2}$. We rewrite it as

$$\left(1-\beta+\sqrt{\varepsilon}\cdot\sqrt{2\beta-\beta^2}\right)\|g\|_D \le (1-\beta+\sqrt{\varepsilon}\cdot\sqrt{2\beta})\|g\|_D \le \left(1-(\sqrt{\beta}-\sqrt{\frac{\varepsilon}{2}})^2+\frac{\varepsilon}{2}\right)\|g\|_D.$$

From all discussion above, $\|\widetilde{f}-f\|_D = \|\alpha(\widetilde{f})-\alpha(f)\|_2 = \|(A^*A)^{-1}\cdot A^*\cdot\vec{y}_w-\alpha(f)\|_2 \le (1+\varepsilon)\|g\|_D$.

### 4.3 Running time of finding ERM

Given the orthonormal basis $v_1,\cdots,v_d$ of $\mathcal{F}$ under $D$, the ERM on noisy observations $y(x_1),\cdots,y(x_m)$ with weights $w_1,\cdots,w_m$ is $(A^*A)^{-1}\cdot A^*\cdot\vec{y}_w$ for $A\in\mathbb{C}^{m\times d}$ defined as $A(i,j)=\sqrt{w_i}\cdot v_j(x_i)$ and $\vec{y}_w = \big(\sqrt{w_1}\cdot y(x_1),\ldots,\sqrt{w_m}\cdot y(x_m)\big)$. Since well-balanced procedures guarantee $\lambda(A^*\cdot A)\in[3/4,5/4]$, we could calculate an $\delta$-approximation of the ERM using Taylor expansion $(A^*A)^{-1}\approx\sum_{i=0}^{t}(I-A^*A)^i$ for $t=O(\log\frac{1}{\delta})$. This saves the cost of calculating the inverse $(A^*A)^{-1}$ and improves it to $O(m\cdot d\cdot\log\frac{1}{\delta})$ for any linear family.

**Observation 4.4.** *Let $A$ be a $m\times d$ matrix defined as $A(i,j)=\sqrt{w_i}\cdot v_j(x_i)$ with $\lambda(A^*\cdot A)\in[3/4,5/4]$. Given $\delta$, for $t=O(\log 1/\delta)$ and any vector $\vec{y}\in\mathbb{R}^m$,*

$$\|(A^*\cdot A)^{-1}\cdot A^*\cdot\vec{y}-\big(\sum_{i=0}^{t}(I-A^*\cdot A)^i\big)\cdot A^*\cdot\vec{y}\|_2 \le \delta\cdot\|(A^*\cdot A)^{-1}\cdot A^*\cdot\vec{y}\|_2.$$

## 5 A Linear-Sample Algorithm for Known $D$

We provide a well-balanced sampling procedure with a linear number of random samples in this section. The procedure requires knowing the underlying distribution $D$, which makes it directly useful in the query setting or the "fixed design" active learning setting, where $D$ can be set to the empirical distribution $D_0$.

**Lemma 5.1.** *Given any dimension $d$ linear space $\mathcal{F}$, any distribution $D$ over the domain of $\mathcal{F}$, and any $\varepsilon>0$, there exists an efficient $\varepsilon$-*well-balanced sampling procedure *that terminates in $O(d/\varepsilon)$ rounds with probability $1-\frac{1}{200}$.*

Theorem 1.1 follows from Theorem 2.3 using the above *well-balanced sampling procedure*. We state the following version for specific types of noise after plugging the *well-balanced sampling procedure* in Lemma 5.1 to Corollary 4.1.

**Theorem 5.2.** *Given any dimension $d$ linear space $\mathcal{F}$ of functions and any distribution $D$ on the domain of $\mathcal{F}$, let $y(x)=f(x)+g(x)$ be our observed function, where $f\in\mathcal{F}$ and $g$ denotes a noise function. For any $\varepsilon>0$, there exists an efficient algorithm that observes $y(x)$ at $m=O(\frac{d}{\varepsilon})$ points and outputs $\widetilde{f}$ such that in expectation,*

1. *$\|\tilde{f}-f\|_D^2\le\varepsilon\cdot\mathop{\mathbb{E}}_{g}[\|g\|_D^2]$, when $g(x)$ is a random function from $G$ to $\mathbb{C}$ where each $g(x)$ is an independent random variable with $\mathop{\mathbb{E}}_{g}[g(x)]=0$.*

2. *$\|\tilde{f}-f\|_D\le(1+\varepsilon)\cdot\|g\|_D$ for any other noise function $g$.*

We show how to extract the coefficients $\alpha_1,\cdots,\alpha_m$ from the randomized BSS algorithm by [LS15] in Algorithm 1. Given $\epsilon$, the linear family $\mathcal{F}$, and the distribution $D$, we fix $\gamma=\sqrt{\epsilon}/C_0$ for a constant $C_0$ and $v_1,\ldots,v_d$ to be an orthonormal basis of $\mathcal{F}$ in this section. For convenience, we use $v(x)$ to denote the vector $\big(v_1(x),\ldots,v_d(x)\big)$.

In the rest of this section, we prove Lemma 5.1 in Section 5.1.

**Algorithm 1** A well-balanced sampling procedure based on Randomized BSS

**procedure** RandomizedSamplingBSS($\mathcal{F}, D, \epsilon$)
  Find an orthonormal basis $v_1, \ldots, v_d$ of $\mathcal{F}$ under $D$;
  Set $\gamma = \sqrt{\epsilon}/C_0$ and $\mathsf{mid} = \frac{4d/\gamma}{1/(1-\gamma)-1/(1+\gamma)}$;
  $j = 0; B_0 = 0$;
  $l_0 = -2d/\gamma; u_0 = 2d/\gamma$;
  **while** $u_{j+1} - l_{j+1} < 8d/\gamma$ **do**;
    $\Phi_j = \mathsf{Tr}(u_j I - B_j)^{-1} + \mathsf{Tr}(B_j - l_j I)^{-1}$; ▷ The potential function at iteration $j$.
    Set the coefficient $\alpha_j = \frac{\gamma}{\Phi_j} \cdot \frac{1}{\mathsf{mid}}$;
    Set the distribution $D_j(x) = D(x) \cdot \Big( v(x)^\top (u_j I - B_j)^{-1} v(x) + v(x)^\top (B_j - l_j I)^{-1} v(x) \Big) / \Phi_j$ for $v(x) = \big(v_1(x), \ldots, v_d(x)\big)$;
    Sample $x_j \sim D_j$ and set a scale $s_j = \frac{\gamma}{\Phi_j} \cdot \frac{D(x)}{D_j(x)}$;
    $B_{j+1} = B_j + s_j \cdot v(x_j) v(x_j)^\top$;
    $u_{j+1} = u_j + \frac{\gamma}{\Phi_j(1-\gamma)}$;   $l_{j+1} = l_j + \frac{\gamma}{\Phi_j(1+\gamma)}$;
    $j = j + 1$;
  **end while**
  $m = j$;
  Assign the weight $w_j = s_j/\mathsf{mid}$ for each $x_j$;
**end procedure**

### 5.1 Proof of Lemma 5.1

We state a few properties of randomized BSS [BSS12, LS15] that will be used in this proof. The first property is that matrices $B_1, \ldots, B_m$ in Procedure RandomizedBSS always have bounded eigenvalues.

**Lemma 5.3** ([BSS12, LS15]). *For any $j \in [m]$, $\lambda(B_j) \in (l_j, u_j)$.*

Lemma 3.6 and 3.7 of [LS15] shows that with high probability, the while loop in Procedure RandomizedSamplingBSS finishes within $O(\frac{d}{\gamma^2})$ iterations and guarantees the last matrix $B_m$ is well-conditioned, i.e., $\frac{\lambda_{\max}(B_m)}{\lambda_{\min}(B_m)} \leq \frac{u_m}{l_m} \leq 1 + O(\gamma)$.

**Lemma 5.4** ([LS15]). *There exists a constant $C$ such that with probability at least $1 - \frac{1}{200}$, Procedure RandomizedSamplingBSS takes at most $m = C \cdot d/\gamma^2$ random points $x_1, \ldots, x_m$ and guarantees that $\frac{u_m}{l_m} \leq 1 + 8\gamma$.*

We first show that $(A^* \cdot A)$ is well-conditioned from the definition of $A$. We prove that our choice of $\mathsf{mid}$ is very close to $\sum_{j=1}^m \frac{\gamma}{\phi_j} = \frac{u_m + l_m}{\frac{1}{1-\gamma} + \frac{1}{1+\gamma}} \approx \frac{u_m + l_m}{2}$.

**Claim 5.5.** *After exiting the while loop in Procedure* RandomizedBSS, *we always have*

1. $u_m - l_m \leq 9d/\gamma$.
2. $(1 - \frac{0.5\gamma^2}{d}) \cdot \sum_{j=1}^m \frac{\gamma}{\phi_j} \leq \mathsf{mid} \leq \sum_{j=1}^m \frac{\gamma}{\phi_j}$.

*Proof.* Let us first bound the last term $\frac{\gamma}{\phi_m}$ in the while loop. Since $u_{m-1} - l_{m-1} < 8d/\gamma$, $\phi_m \geq 2d \cdot \frac{1}{4d/\gamma} \geq \frac{\gamma}{2}$, which indicates the last term $\frac{\gamma}{\phi_m} \leq 2$. Thus

$$u_m - l_m \leq 8d/\gamma + 2(\frac{1}{1-\gamma} - \frac{1}{1+\gamma}) \leq 8d/\gamma + 5\gamma.$$

From our choice $\mathsf{mid} = \frac{4d/\gamma}{1/(1-\gamma)-1/(1+\gamma)} = 2d(1-\gamma^2)/\gamma^2$ and the condition of the while loop $u_m - l_m = \sum_{j=1}^{m}(\gamma/\phi_j)\cdot(\frac{1}{1-\gamma}-\frac{1}{1+\gamma}) + 4d/\gamma \geq 8d/\gamma$, we know

$$\sum_{j=1}^{m}\frac{\gamma}{\phi_j} \geq \mathsf{mid} = 2d(1-\gamma^2)/\gamma^2.$$

On the other hand, since $u_{m-1} - l_{m-1} < 8d/\gamma$ is in the while loop, $\sum_{j=1}^{m-1}\frac{\gamma}{\phi_j} < \mathsf{mid}$. Hence

$$\mathsf{mid} > \sum_{j=1}^{m-1}\frac{\gamma}{\phi_j} \geq \sum_{j=1}^{m}\frac{\gamma}{\phi_j} - 2 \geq (1-0.5\gamma^2/d)\cdot(\sum_{j=1}^{m}\frac{\gamma}{\phi_j}).$$

□

**Lemma 5.6.** *Given $\frac{u_m}{l_m} \leq 1+8\gamma$, $\lambda(A^*\cdot A) \in (1-5\gamma, 1+5\gamma)$.*

*Proof.* For $B_m = \sum_{j=1}^{m} s_j v(x_j)v(x_j)^\top$, $\lambda(B_m) \in (l_m, u_m)$ from Lemma 5.3. At the same time, given $w_j = s_j/\mathsf{mid}$,

$$(A^*A) = \sum_{j=1}^{m} w_j v(x_j)v(x_j)^\top = \frac{1}{\mathsf{mid}}\cdot\sum_{j=1}^{m} s_j v(x_j)v(x_j)^\top = \frac{B_m}{\mathsf{mid}}.$$

Since $\mathsf{mid} \in [1-\frac{3\gamma^2}{d}, 1]\cdot(\sum_{j=1}^{m}\frac{\gamma}{\phi_j}) = [1-\frac{3\gamma^2}{d}, 1]\cdot(\frac{u_m+l_m}{\frac{1}{1-\gamma}+\frac{1}{1+\gamma}}) \subseteq [1-2\gamma^2, 1-\gamma^2]\cdot(\frac{u_m+l_m}{2})$ from Claim 5.5, $\lambda(A^*\cdot A) = \lambda(B_m)/\mathsf{mid} \in (l_m/\mathsf{mid}, u_m/\mathsf{mid}) \subset (1-5\gamma, 1+5\gamma)$ given $\frac{u_m}{l_m} \leq 1+8\gamma$ in Lemma 5.4. □

We finish the proof of Lemma 5.1 by combining all discussion above.

*Proof of Lemma 5.1.* From Lemma 5.4 and Lemma 5.6, $m = O(d/\gamma^2)$ and $\lambda(A^*A) \in [1-1/4, 1+1/4]$ with probability 0.995.

For $\alpha_i = \frac{\gamma}{\Phi_i}\cdot\frac{1}{\mathsf{mid}}$, we bound $\sum_{i=1}^{m}\frac{\gamma}{\Phi_i}\cdot\frac{1}{\mathsf{mid}}$ by 1.25 from the second property of Claim 5.5.

Then we bound $\alpha_j \cdot K_{D_j}$. We notice that $\sup_{h\in\mathcal{F}}\frac{|h(x)|^2}{\|h\|_D^2} = \sum_{i\in[d]}|v_i(x)|^2$ for every $x \in G$ because $\sup_{h\in\mathcal{F}}\frac{|h(x)|^2}{\|h\|_D^2} = \sup_{\alpha(h)}\frac{\left|\sum_i \alpha(h)_i\cdot v_i(x)\right|^2}{\|\alpha(h)\|_2^2} = \sum_i |v_i(x)|^2$ by the Cauchy-Schwartz inequality. This simplifies $K_{D_j}$ to $\sup_x\{\frac{D(x)}{D_j(x)}\cdot\sum_{i=1}^{d}|u_i(x)|^2\}$ and bounds $\alpha_j\cdot K_{D_j}$ by

$$\begin{aligned}
&\frac{\gamma}{\Phi_j\cdot\mathsf{mid}}\cdot\sup_x\left\{\frac{D(x)}{D_j(x)}\cdot\sum_{i=1}^{d}|v_i(x)|^2\right\}\\
=&\frac{\gamma}{\mathsf{mid}}\cdot\sup_x\left\{\frac{\sum_{i=1}^{d}|v_i(x)|^2}{v(x_j)^\top(u_jI-B_j)^{-1}v(x_j)+v(x_j)^\top(B_j-l_jI)^{-1}v(x_j)}\right\}\\
\leq&\frac{\gamma}{\mathsf{mid}}\cdot\sup_x\left\{\frac{\sum_{i=1}^{d}|v_i(x)|^2}{\lambda_{\min}\big((u_jI-B_j)^{-1}\big)\cdot\|v(x_j)\|_2^2+\lambda_{\min}\big((B_j-l_jI)^{-1}\big)\cdot\|v(x_j)\|_2^2}\right\}\\
\leq&\frac{\gamma}{\mathsf{mid}}\cdot\frac{1}{1/(u_j-l_j)+1/(u_j-l_j)}\\
=&\frac{\gamma}{\mathsf{mid}}\cdot\frac{u_j-l_j}{2} \qquad \text{(apply the first property of Claim 5.5)}\\
\leq&\frac{4.5\cdot d}{\mathsf{mid}} \leq 3\gamma^2 = 3\epsilon/C_0^2.
\end{aligned}$$

By choosing $C_0 = 3$, this satisfies the second property of *well-balanced sampling procedures*. At the same time, by Lemma 4.2, Algorithm 1 also satisfies the first property of *well-balanced sampling procedures*. $\square$

# 6 Performance of i.i.d. Distributions

Given the linear family $\mathcal{F}$ of dimension $d$ and the measure of distance $D$, we provide a distribution $D_{\mathcal{F}}$ with a condition number $K_{D_{\mathcal{F}}} = d$.

**Lemma 6.1.** *Given any linear family $\mathcal{F}$ of dimension $d$ and any distribution $D$, there always exists an explicit distribution $D_{\mathcal{F}}$ such that the condition number*

$$K_{D_{\mathcal{F}}} = \sup_x \left\{ \sup_{h \in \mathcal{F}} \left\{ \frac{D(x)}{D_{\mathcal{F}}(x)} \cdot \frac{|h(x)|^2}{\|h\|_D^2} \right\} \right\} = d.$$

Next, for generality, we bound the number of i.i.d. random samples from an arbitrary distribution $D'$ to fulfill the requirements of *well-balanced sampling procedures* in Definition 2.1.

**Lemma 6.2.** *There exists a universal constant $C_1$ such that given any distribution $D'$ with the same support of $D$ and any $\epsilon > 0$, the random sampling procedure with $m = C_1(K_{D'} \log d + \frac{K_{D'}}{\varepsilon})$ i.i.d. random samples from $D'$ and coefficients $\alpha_1 = \cdots = \alpha_m = 1/m$ is an $\varepsilon$*-well-balanced sampling procedure.

By Theorem 2.3, we state the following result, which will be used in active learning. For $G = \mathsf{supp}(D)$ and any $x \in G$, let $Y(x)$ denote the conditional distribution $(Y|D = x)$ and $(D', Y(D'))$ denote the distribution that first generates $x \sim D'$ then generates $y \sim Y(x)$.

**Theorem 6.3.** *Consider any dimension $d$ linear space $\mathcal{F}$ of functions from a domain $G$ to $\mathbb{C}$. Let $(D, Y)$ be a joint distribution over $G \times \mathbb{C}$, and $f = \arg\min_{h \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y - h(x)|^2]$.*

*Let $D'$ be any distribution on $G$ and $K_{D'} = \sup_x \left\{ \sup_{h \in \mathcal{F}} \left\{ \frac{D(x)}{D'(x)} \cdot \frac{|h(x)|^2}{\|h\|_D^2} \right\} \right\}$. The weighted ERM $\widetilde{f}$ of $m = O(K_{D'} \log d + \frac{K_{D'}}{\varepsilon})$ random queries of $(D', Y(D'))$ with weights $w_i = \frac{D(x_i)}{m \cdot D'(x_i)}$ for each $i \in [m]$ satisfies*

$$\|\tilde{f} - f\|_D^2 = \mathop{\mathbb{E}}_{x \sim D}\left[|\tilde{f}(x) - f(x)|^2\right] \leq \varepsilon \cdot \mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}\left[|y - f(x)|^2\right] \text{ with probability } 1 - 10^{-4}.$$

We show the proof of Lemma 6.1 in Section 6.1 and the proof of Lemma 6.2 in Section 6.2.

## 6.1 Optimal Condition Number

We describe the distribution $D_{\mathcal{F}}$ with $K_{D_{\mathcal{F}}} = d$. We first observe that for any family $\mathcal{F}$ (not necessarily linear), we could always scale down the condition number to $\kappa = \mathop{\mathbb{E}}_{x \sim D}\left[\sup_{h \in \mathcal{F}: h \neq 0} \frac{|h(x)|^2}{\|h\|_D^2}\right]$.

**Claim 6.4.** *For any family $\mathcal{F}$ and any distribution $D$ on its domain, let $D_{\mathcal{F}}$ be the distribution defined as $D_{\mathcal{F}}(x) = \frac{D(x) \cdot \sup_{h \in \mathcal{F}: h \neq 0} \frac{|h(x)|^2}{\|h\|_D^2}}{\kappa}$ with $\kappa$. The condition number $K_{D_{\mathcal{F}}}$ is at most $\kappa$.*

**Algorithm 2** SampleDF

1: **procedure** GeneratingDF($\mathcal{F} = \text{span}\{v_1, \ldots, v_d\}, D$)
2: Sample $j \in [d]$ uniformly.
3: Sample $x$ from the distribution $W_j(x) = D(x) \cdot |v_j(x)|^2$.
4: Set the weight of $x$ to be $\frac{d}{\sum_{i=1}^d |v_i(x)|^2}$.
5: **end procedure**

*Proof.* For any $g \in \mathcal{F}$ and $x$ in the domain $G$,

$$\frac{|g(x)|^2}{\|g\|_D^2} \cdot \frac{D(x)}{D_{\mathcal{F}}(x)} = \frac{\frac{|g(x)|^2}{\|g\|_D^2} \cdot D(x)}{\sup_{h \in \mathcal{F}} \frac{|h(x)|^2}{\|h\|_D^2} \cdot D(x)/\kappa} \leq \kappa.$$

□

Next we use the linearity of $\mathcal{F}$ to prove $\kappa = d$. Let $\{v_1, \ldots, v_d\}$ be any orthonormal basis of $\mathcal{F}$, where inner products are taken under the distribution $D$.

**Lemma 6.5.** *For any linear family $\mathcal{F}$ of dimension $d$ and any distribution $D$,*

$$\mathop{\mathbb{E}}_{x \sim D} \sup_{h \in \mathcal{F}: \|h\|_D = 1} |h(x)|^2 = d$$

*such that $D_{\mathcal{F}}(x) = D(x) \cdot \sup\limits_{h \in \mathcal{F}: \|h\|_D = 1} |h(x)|^2/d$ has a condition number $K_{D_{\mathcal{F}}} = d$. Moreover, there exists an efficient algorithm to sample $x$ from $D_{\mathcal{F}}$ and compute its weight $\frac{D(x)}{D_{\mathcal{F}}(x)}$.*

*Proof.* Given an orthonormal basis $v_1, \ldots, v_d$ of $\mathcal{F}$, for any $h \in \mathcal{F}$ with $\|h\|_D = 1$, there exists $c_1, \ldots, c_d$ such that $h(x) = c_i \cdot v_i(x)$. Then for any $x$ in the domain, from the Cauchy-Schwartz inequality,

$$\sup_h \frac{|h(x)|^2}{\|h\|_D^2} = \sup_{c_1, \ldots, c_d} \frac{|\sum_{i \in [d]} c_i v_i(x)|^2}{\sum_{i \in [d]} |c_i|^2} = \frac{(\sum_{i \in [d]} |c_i|^2) \cdot (\sum_{i \in [d]} |v_i(x)|^2)}{\sum_{i \in [d]} |c_i|^2} = \sum_{i \in [d]} |v_i(x)|^2.$$

This is tight because there always exist $c_1 = \overline{v_1(x)}, c_2 = \overline{v_2(x)}, \ldots, c_d = \overline{v_d(x)}$ such that $|\sum\limits_{i \in [d]} c_i v_i(x)|^2 = (\sum\limits_{i \in [d]} |c_i|^2) \cdot (\sum\limits_{i \in [d]} |v_i(x)|^2)$. Hence

$$\mathop{\mathbb{E}}_{x \sim D} \sup_{h \in \mathcal{F}: h \neq 0} \frac{|h(x)|^2}{\|h\|_D^2} = \mathop{\mathbb{E}}_{x \sim D} \Big[ \sum_{i \in [d]} |v_i(x)|^2 \Big] = d.$$

By Claim 6.4, this indicates $K_{D_{\mathcal{F}}} = d$. At the same time, this calculation indicates

$$D_{\mathcal{F}}(x) = \frac{D(x) \cdot \sup\limits_{\|h\|_D = 1} |h(x)|^2}{d} = \frac{D(x) \cdot \sum_{i \in [d]} |v_i(x)|^2}{d}.$$

We present our sampling procedure in Algorithm 2.

□

## 6.2 Proof of Lemma 6.2

We use the matrix Chernoff theorem to prove the first property in Definition 2.1. We still use $A$ to denote the $m \times d$ matrix $A(i,j) = \sqrt{w_i} \cdot v_j(x_i)$.

**Lemma 6.6.** *Let $D'$ be an arbitrary distribution over $G$ and*

$$K_{D'} = \sup_{h \in \mathcal{F}: h \neq 0} \sup_{x \in G} \frac{|h^{(D')}(x)|^2}{\|h\|_D^2}. \tag{8}$$

*There exists an absolute constant $C$ such that for any $n \in \mathbb{N}^+$, linear family $\mathcal{F}$ of dimension $d$, $\varepsilon \in (0,1)$ and $\delta \in (0,1)$, when $S = (x_1, \dots, x_m)$ are independently from the distribution $D'$ with $m \geq \frac{C}{\varepsilon^2} \cdot K_{D'} \log \frac{d}{\delta}$ and $w_j = \frac{D(x_j)}{m \cdot D'(x_j)}$ for each $j \in [m]$, the $m \times d$ matrix $A(i,j) = \sqrt{w_i} \cdot v_j(x_i)$ satisfies*

$$\|A^* A - I\| \leq \varepsilon \text{ with probability at least } 1 - \delta.$$

Before we prove Lemma 6.6, we state the following version of the matrix Chernoff bound.

**Theorem 6.7** (Theorem 1.1 of [Tro12])**.** *Consider a finite sequence $\{X_k\}$ of independent, random, self-adjoint matrices of dimension $d$. Assume that each random matrix satisfies*

$$X_k \succeq 0 \quad \text{and} \quad \lambda(X_k) \leq R.$$

*Define $\mu_{\min} = \lambda_{\min}(\sum_k \mathbb{E}[X_k])$ and $\mu_{\max} = \lambda_{\max}(\sum_k \mathbb{E}[X_k])$. Then*

$$\Pr\left\{\lambda_{\min}(\sum_k X_k) \leq (1-\delta)\mu_{\min}\right\} \leq d\left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu_{\min}/R} \quad \text{for } \delta \in [0,1], \text{and} \tag{9}$$

$$\Pr\left\{\lambda_{\max}(\sum_k X_k) \geq (1+\delta)\mu_{\max}\right\} \leq d\left(\frac{e^{-\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu_{\max}/R} \quad \text{for } \delta \geq 0 \tag{10}$$

*Proof.* Let $v_1, \dots, v_d$ be the orthonormal basis of $\mathcal{F}$ in the definition of matrix $A$. For any $h \in \mathcal{F}$, let $\alpha(h) = (\alpha_1, \dots, \alpha_d)$ denote the coefficients of $h$ under $v_1, \dots, v_d$ such that $\|h\|_D^2 = \|\alpha(h)\|_2^2$. At the same time, for any fixed $x$, $\sup\limits_{h \in \mathcal{F}} \frac{|h^{(D')}(x)|^2}{\|h\|_D^2} = \sup\limits_{\alpha(h)} \frac{|\sum_{i=1}^d \alpha(h)_i \cdot v_i^{(D')}(x)|^2}{\|\alpha(h)\|_2^2} = \sum_{i \in [d]} |v_i^{(D')}(x)|^2$ by the tightness of the Cauchy Schwartz inequality. Thus

$$K_{D'} \overset{def}{=} \sup_{x \in G}\{\sup_{h \in \mathcal{F}: h \neq 0} \frac{|h^{(D')}(x)|^2}{\|h\|_D^2}\} \quad \text{indicates} \quad \sup_{x \in G} \sum_{i \in [d]} |v_i^{(D')}(x)|^2 \leq K_{D'}. \tag{11}$$

For each point $x_j$ in $S$ with weight $w_j = \frac{D(x_j)}{m \cdot D'(x_j)}$, let $A_j$ denote the $j$th row of the matrix $A$. It is a vector in $\mathbb{C}^d$ defined by $A_j(i) = A(j,i) = \sqrt{w_j} \cdot v_i(x_j) = \frac{v_i^{(D')}(x_j)}{\sqrt{m}}$. So $A^* A = \sum_{j=1}^m A_j^* \cdot A_j$.

For $A_j^* \cdot A_j$, it is always $\succeq 0$. Notice that the only non-zero eigenvalue of $A_j^* \cdot A_j$ is

$$\lambda(A_j^* \cdot A_j) = A_j \cdot A_j^* = \frac{1}{m}\left(\sum_{i \in [d]} |v_i^{(D')}(x_j)|^2\right) \leq \frac{K_{D'}}{m}$$

from (11).

At the same time, $\sum_{j=1}^{m} \mathbb{E}[A_j^* \cdot A_j]$ equals the identity matrix of size $d \times d$ because the expectation of the entry $(i, i')$ in $A_j^* \cdot A_j$ is

$$\begin{aligned}\mathop{\mathbb{E}}_{x_j \sim D'}[\overline{A(j,i)} \cdot A(j,i')] &= \mathop{\mathbb{E}}_{x_j \sim D'}[\frac{\overline{v_i^{(D')}(x_j)} \cdot v_{i'}^{(D')}(x_j)}{m}] \\ &= \mathop{\mathbb{E}}_{x_j \sim D'}[\frac{D(x) \cdot \overline{v_i(x_j)} \cdot v_{i'}(x_j)}{m \cdot D'(x_j)}] = \mathop{\mathbb{E}}_{x_j \sim D}[\frac{\overline{v_i(x_j)} \cdot v_{i'}(x_j)}{m}] = 1_{\vec{i}=\vec{i'}}/m.\end{aligned}$$

Now we apply Theorem 6.7 on $A^* A = \sum_{j=1}^{m} (A_j^* \cdot A_j)$:

$$\begin{aligned}\Pr\left[\lambda(A^* A) \notin [1-\varepsilon, 1+\varepsilon]\right] &\leq d \left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right)^{1/\frac{K_{D'}}{m}} + d \left(\frac{e^{-\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{1/\frac{K_{D'}}{m}} \\ &\leq 2d \cdot e^{-\frac{\varepsilon^2 \cdot \frac{m}{K_{D'}}}{3}} \leq \delta \qquad\qquad \text{given } m \geq \frac{6 K_{D'} \log \frac{d}{\delta}}{\varepsilon^2}.\end{aligned}$$

□

Then we finish the proof of Lemma 6.2.

*Proof of Lemma 6.2.* Because the coefficient $\alpha_i = 1/m = O(\varepsilon/K_{D'})$ and $\sum_i \alpha_i = 1$, this indicates the second property of *well-balanced sampling procedures*.

Since $m = \Theta(K_{D'} \log d)$, by Lemma 6.6, we know all eigenvalues of $A^* \cdot A$ are in $[1-1/4, 1+1/4]$ with probability $1 - 10^{-3}$. By Lemma 4.2, this indicates the first property of *well-balanced sampling procedures*. □

# 7 Results for Active Learning

In this section, we investigate the case where we do not know the distribution $D$ of $x$ and only receive random samples from $D$. We finish the proof of Theorem 2.4 that bounds the number of unlabeled samples by the condition number of $D$ and the number of labeled samples by $dim(\mathcal{F})$ to find the truth through $D$.

**Theorem 2.4.** *Consider any dimension $d$ linear space $\mathcal{F}$ of functions from a domain $G$ to $\mathbb{C}$. Let $(D, Y)$ be a joint distribution over $G \times \mathbb{C}$ and $f = \arg\min_{h \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y) \sim (D,Y)}[|y - h(x)|^2]$.*

*Let $K = \sup_{h \in \mathcal{F}: h \neq 0} \frac{\sup_{x \in G} |h(x)|^2}{\|h\|_D^2}$. For any $\varepsilon > 0$, there exists an efficient algorithm that takes $O(K \log d + \frac{K}{\varepsilon})$ unlabeled samples from $D$ and requests $O(\frac{d}{\varepsilon})$ labels to output $\widetilde{f}$ satisfying*

$$\mathop{\mathbb{E}}_{x \sim D}[|\tilde{f}(x) - f(x)|^2] \leq \varepsilon \cdot \mathop{\mathbb{E}}_{(x,y) \sim (D,Y)}[|y - f(x)|^2] \text{ in expectation.}$$

Notice that Theorem 1.2 follows from Corollary 4.1 and the guarantee of Theorem 2.4. For generality, we bound the number of labels using any *well-balanced sampling procedure*, such that Theorem 2.4 follows from this lemma with the linear sample procedure in Lemma 5.1.

**Lemma 7.1.** *Consider any dimension $d$ linear space $\mathcal{F}$ of functions from a domain $G$ to $\mathbb{C}$. Let $(D, Y)$ be a joint distribution over $G \times \mathbb{C}$ and $f = \arg\min_{h \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y) \sim (D,Y)}[|y - h(x)|^2]$.*

*Let* $K = \sup_{h \in \mathcal{F}: h \neq 0} \frac{\sup_{x \in G} |h(x)|^2}{\|h\|_D^2}$ *and* $P$ *be a* well-balanced sampling procedure *terminating in* $m_p(\varepsilon)$ *rounds with probability* $1 - 10^{-3}$ *for any linear family* $\mathcal{F}$*, measurement* $D$*, and* $\varepsilon$*. For any* $\varepsilon > 0$*, Algorithm 3 takes* $O(K \log d + \frac{K}{\varepsilon})$ *unlabeled samples from* $D$ *and requests at most* $m_p(\varepsilon/8)$ *labels to output* $\widetilde{f}$ *satisfying*

$$\mathop{\mathbb{E}}_{x \sim D}[|\tilde{f}(x) - f(x)|^2] \leq \varepsilon \cdot \mathop{\mathbb{E}}_{(x,y) \sim (D,Y)}[|y - f(x)|^2] \text{ in expectation.}$$

Algorithm 3 first takes $m_0 = O(K \log d + K/\varepsilon)$ unlabeled samples and defines a distribution $D_0$ to be the uniform distribution on these $m_0$ samples. Then it uses $D_0$ to simulate $D$ in $P$, i.e., it outputs the ERM of a good execution of the well-balanced sampling procedure $P$ with the linear family $\mathcal{F}$, the measurement $D_0$, and $\frac{\epsilon}{8}$.

---

**Algorithm 3** Regression over an unknown distribution $D$

---

1: **procedure** RegressionUnknownDistribution($\varepsilon, \mathcal{F}, D, P$)
2: Set $C$ to be a large constant and $m_0 = C \cdot (K \log d + K/\varepsilon)$ .
3: Take $m_0$ unlabeled samples $x_1, \ldots, x_{m_0}$ from $D$.
4: Let $D_0$ be the uniform distribution over $(x_1, \ldots, x_{m_0})$.
5: Output the ERM $\tilde{f}$ of a good execution of $P$ with parameters $\mathcal{F}, D_0, \epsilon/8$.
6: **end procedure**

---

*Proof.* We still use $\|f\|_{D'}$ to denote $\sqrt{\mathop{\mathbb{E}}_{x \sim D'}[|f(x)|^2]}$ and $D_1$ to denote the weighted distribution generated by Procedure $P$ given $\mathcal{F}, D_0, \varepsilon$. By Lemma 6.2 with $D$ and the property of $P$, with probability at least $1 - 2 \cdot 10^{-3}$,

$$\|h\|_{D_0}^2 = (1 \pm 1/4) \cdot \|h\|_D^2 \text{ and } \|h\|_{D_1}^2 = (1 \pm 1/4) \cdot \|h\|_{D_0}^2 \text{ for every } h \in \mathcal{F}. \tag{12}$$

We assume (12) holds in the rest of this proof.

Let $y_i$ denote a random label of $x_i$ from $Y(x_i)$ for each $i \in [m_0]$ including the unlabeled samples in the algorithm and the labeled samples in Step 5 of Algorithm 3. Let $f'$ be the weighted ERM of $(x_1, \cdots, x_m)$ and $(y_1, \cdots, y_m)$ over $D_0$, i.e.,

$$f' = \arg\min_{h \in \mathcal{F}} \mathop{\mathbb{E}}_{x_i \sim D_0, y_i \sim Y(x_i)} \left[|y_i - h(x_i)|^2\right]. \tag{13}$$

Given Property (12) and Lemma 6.2,

$$\mathop{\mathbb{E}}_{(x_1,y_1),\ldots,(x_{m_0},y_{m_0})}[\|f' - f\|_D^2] \leq \varepsilon \cdot \mathop{\mathbb{E}}_{(x,y) \sim (D,Y)}[|y - f(x)|^2] \text{ from the proof of Theorem 2.3 .}$$

In the rest of this proof, we show that the weighted ERM $\widetilde{f}$ of a good execution of $P$ with measurement $D_0$ guarantees $\|\widetilde{f} - f'\|_{D_0}^2 \lesssim \mathop{\mathbb{E}}_{(x,y) \sim (D,Y)}\left[|y - f(x)|^2\right]$ with high probability. Given Property (12) and the guarantee of Procedure $P$, we have

$$\mathop{\mathbb{E}}_{P}[\|\widetilde{f} - f'\|_{D_0}^2] \leq \varepsilon \cdot \mathop{\mathbb{E}}_{x \sim D_0}\left[|y_i - f'(x_i)|^2\right]$$

from the proof of Theorem 2.3. Next we bound the right hand side $\mathop{\mathbb{E}}_{x_i \sim D_0}\left[|y_i - f'(x_i)|^2\right]$ by

$\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}\left[|y-f(x)|^2\right]$ over the randomness of $(x_1,y_1),\ldots,(x_{m_0},y_{m_0})$:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{(x_1,y_1),\ldots,(x_{m_0},y_{m_0})}\left[\mathop{\mathbb{E}}_{x_i\sim D_0}\left[|y_i-f'(x_i)|^2\right]\right]\\
\le&\mathop{\mathbb{E}}_{(x_1,y_1),\ldots,(x_{m_0},y_{m_0})}\left[2\mathop{\mathbb{E}}_{x_i\sim D_0}\left[|y_i-f(x_i)|^2\right]+2\|f-f'\|_{D_0}^2\right]\\
\le&2\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}\left[|y-f(x)|^2\right]+3\mathop{\mathbb{E}}_{(x_1,y_1),\ldots,(x_{m_0},y_{m_0})}\left[\|f-f'\|_D^2\right] \quad \text{from (12)}
\end{aligned}$$

Hence $\mathop{\mathbb{E}}_{(x_1,y_1),\ldots,(x_{m_0},y_{m_0})}\left[\mathop{\mathbb{E}}_P[\|\widetilde{f}-f'\|_{D_0}^2]\right]\lesssim\varepsilon\cdot\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}\left[|y-f(x)|^2\right]$.

From all discussion above, by rescaling $\varepsilon$, we have

$$\|\widetilde{f}-f\|_D^2\le2\|\widetilde{f}-f'\|_D^2+2\|f'-f\|_D^2\le3\|\widetilde{f}-f'\|_{D_0}^2+\frac{\varepsilon}{4}\cdot\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y-f(x)|^2]\le\varepsilon\cdot\mathop{\mathbb{E}}_{(x,y)\sim(D,Y)}[|y-f(x)|^2]$$

□

# 8 Lower Bounds

We present two lower bounds on the number of samples in this section. We first prove a lower bound on the query complexity based on the dimension $d$. Then we prove a lower bound on the the sample complexity based on the condition number of the sampling distribution.

**Theorem 8.1.** *For any $d$ and any $\varepsilon<\frac{1}{10}$, there exist a distribution $D$ and a linear family $\mathcal{F}$ of functions with dimension $d$ such that for the i.i.d. Gaussian noise $g(x)=N(0,\frac{1}{\varepsilon})$, any algorithm which observes $y(x)=f(x)+g(x)$ for $f\in\mathcal{F}$ with $\|f\|_D=1$ and outputs $\widetilde{f}$ satisfying $\|f-\widetilde{f}\|_D\le0.1$ with probability $\ge\frac{3}{4}$, needs at least $m\ge\frac{0.8d}{\varepsilon}$ queries.*

Notice that this lower bound matches the upper bound in Theorem 1.1 up to a constant factor. In the rest of this section, we focus on the proof of Theorem 8.1. Let $\mathcal{F}=\{f:[d]\to\mathbb{R}\}$ and $D$ be the uniform distribution over $[d]$. We first construct a packing set $\mathcal{M}$ of $\mathcal{F}$.

**Claim 8.2.** *There exists a subset $\mathcal{M}=\{f_1,\ldots,f_n\}\subseteq\mathcal{F}$ with the following properties:*

1. *$\|f_i\|_D=1$ for each $f_i\in\mathcal{M}$.*
2. *$\|f_i\|_\infty\le1$ for each $f_i\in\mathcal{M}$.*
3. *$\|f_i-f_j\|_D>0.2$ for distinct $f_i,f_j$ in $\mathcal{M}$.*
4. *$n\ge2^{0.7d}$.*

*Proof.* We construct $\mathcal{M}$ from $U=\left\{f:[d]\to\{\pm1\}\right\}$ in Procedure ConstructM. Notice that $|U|=2^d$ before the while loop. At the same time, Procedure ConstructM removes at most $\binom{d}{\le0.01d}\le2^{0.3d}$ functions every time because $\|g-h\|_D<0.2$ indicates $\Pr[g(x)\ne h(x)]\le(0.2)^2/4=0.01$. Thus $n\ge2^d/2^{0.3d}\ge2^{0.7d}$.

**Algorithm 4** Construct $\mathcal{M}$

```
procedure ConstructM(d)
    Set n = 0 and U = {f : [d] → {±1}}.
    while U ≠ ∅ do
        Choose any h ∈ U and remove all functions h' ∈ U with ‖h − h'‖_D < 0.2.
        n = n + 1 and f_n = h.
    end while
    Return M = {f_1, ..., f_n}.
end procedure
```

□

We finish the proof of Theorem 8.1 using the Shannon-Hartley theorem.

**Theorem 8.3** (The Shannon-Hartley Theorem [Har28, Sha49]). *Let $S$ be a real-valued random variable with $\mathbb{E}[S^2] = \tau^2$ and $T \sim N(0, \sigma^2)$. The mutual information $I(S; S+T) \leq \frac{1}{2}\log(1 + \frac{\tau^2}{\sigma^2})$.*

*Proof of Theorem 8.1.* Because of Yao's minimax principle, we assume $A$ is a deterministic algorithm given the i.i.d. Gaussian noise. Let $I(\widetilde{f}; f_j)$ denote the mutual information of a random function $f_j \in \mathcal{M}$ and $A$'s output $\widetilde{f}$ given $m$ observations $(x_1, y_1), \ldots, (x_m, y_m)$ with $y_i = f_j(x_i) + N(0, \frac{1}{\varepsilon})$. When the output $\widetilde{f}$ satisfies $\|\widetilde{f} - f_j\|_D \leq 0.1$, $f_j$ is the closest function to $\widetilde{f}$ in $\mathcal{M}$ from the third property of $\mathcal{M}$. From Fano's inequality [Fan61], $H(f_j|\tilde{f}) \leq H(\frac{1}{4}) + \frac{\log(|\mathcal{M}|-1)}{4}$. This indicates

$$I(f_j; \widetilde{f}) = H(f_j) - H(f_j|\tilde{f}) \geq \log|\mathcal{M}| - 1 - \log(|\mathcal{M}|-1)/4 \geq 0.7\log|\mathcal{M}| \geq 0.4d.$$

At the same time, by the data processing inequality, the algorithm $A$ makes $m$ queries $\big(x_1, \ldots, x_m\big)$ and sees $\big(y_1, \ldots, y_m\big)$, which indicates

$$I(\widetilde{f}; f_j) \leq I\Big(\big(y_1, \ldots, y_m\big); f_j\Big) = \sum_{i=1}^{m} I\Big(y_i; f_j(x_i)\big|y_1, \cdots, y_{i-1}\Big). \tag{14}$$

For the query $x_i$, let $D_{i,j}$ denote the distribution of $f_j \in \mathcal{M}$ in the algorithm $A$ given the first $i-1$ observations $\big(x_1, y_1\big), \ldots, \big(x_{i-1}, y_{i-1}\big)$. We apply Theorem 8.3 on $D_{i,j}$ such that it bounds

$$\begin{aligned} I\left(y_i = f_j(x_i) + N(0, \frac{1}{\varepsilon}); f_j(x_i)\big|y_1, \cdots, y_{i-1}\right) \leq& \frac{1}{2}\log\left(1 + \frac{\mathbb{E}_{f \sim D_{i,j}}[f(x_i)^2]}{1/\epsilon}\right) \\ \leq& \frac{1}{2}\log\left(1 + \frac{\max_{f \in \mathcal{M}}[f(x_i)^2]}{1/\varepsilon}\right) \\ =& \frac{1}{2}\log\big(1 + \varepsilon\big) \leq \frac{\varepsilon}{2}, \end{aligned}$$

where we apply the second property of $\mathcal{M}$ in the second step to bound $f(x)^2$ for any $f \in \mathcal{M}$. Hence we bound $\sum_{i=1}^{m} I(y_i; f_j|y_1, \cdots, y_{i-1})$ by $m \cdot \frac{\varepsilon}{2}$. This implies

$$0.4d \leq m \cdot \frac{\varepsilon}{2} \Rightarrow m \geq \frac{0.8d}{\varepsilon}.$$

$\square$

Next we consider the sample complexity of linear regression.

**Theorem 8.4.** *For any $K$, $d$, and $\varepsilon > 0$, there exist a distribution $D$, a linear family of functions $\mathcal{F}$ with dimension $d$ whose condition number $\sup_{h \in \mathcal{F}: h \neq 0} \left\{ \sup_{x \in G} \frac{|h(x)|^2}{\|h\|_D^2} \right\}$ equals $K$, and a noise function $g$ orthogonal to $V$ such that any algorithm observing $y(x) = f(x) + g(x)$ of $f \in \mathcal{F}$ needs at least $\Omega(K \log d + \frac{K}{\varepsilon})$ samples from $D$ to output $\widetilde{f}$ satisfying $\|\widetilde{f} - f\|_D \leq 0.1\sqrt{\varepsilon} \cdot \|g\|_D$ with probability $\frac{3}{4}$.*

*Proof.* We fix $K$ to be an integer and set the domain of functions in $\mathcal{F}$ to be $[K]$. We choose $D$ to be the uniform distribution over $[K]$. Let $\mathcal{F}$ denote the family of functions $\{f : [K] \to \mathbb{C} | f(d+1) = f(d+2) = \cdots = f(K) = 0\}$. Its condition number $\sup_{h \in \mathcal{F}: h \neq 0} \{\sup_{x \in G} \frac{|h(x)|^2}{\|h\|_D^2}\}$ equals $K$. $h(x) = 1_{x=1}$ provides the lower bound $\geq K$. At the same time, $\frac{|h(x)|^2}{\|h\|_D^2} = \frac{|h(x)|^2}{\sum_{i=1}^K |h(x)|^2/K} \leq K$ indicates the upper bound $\leq K$.

We first consider the case $K \log d \geq \frac{K}{\varepsilon}$. Let $g = 0$ such that $g$ is orthogonal to $V$. Notice that $\|\widetilde{f} - f\|_D \leq 0.1\sqrt{\varepsilon} \cdot \|g\|_D$ indicates $\widetilde{f}(x) = f(x)$ for every $x \in [d]$. Hence the algorithm needs to sample $f(x)$ for every $x \in [d]$ when sampling from $D$: the uniform distribution over $[K]$. From the lower bound of the coupon collector problem, this takes at least $\Omega(K \log d)$ samples from $D$.

Otherwise, we prove that the algorithm needs $\Omega(K/\varepsilon)$ samples. Without loss of generality, we assume $\mathop{\mathbb{E}}_{x \sim [d]}\left[|f(x)|^2\right] = 1$ for the truth $f$ in $y$. Let $g(x) = N(0, 1/\varepsilon)$ for each $x \in [d]$. From Theorem 8.1, to find $\widetilde{f}$ satisfying $\mathop{\mathbb{E}}_{x \sim [d]}\left[|\widetilde{f}(x) - f(x)|^2\right] \leq 0.1 \mathop{\mathbb{E}}_{x \sim [d]}\left[|f(x)|^2\right]$, the algorithm needs at least $\Omega(d/\varepsilon)$ queries of $x \in [d]$. Hence it needs $\Omega(K/\varepsilon)$ random samples from $D$, the uniform distribution over $[K]$. $\square$

# 9 Application to Continuous $k$-sparse Fourier Transforms

We consider the nonlinear function space containing signals with $k$-sparse Fourier transform in the continuous setting. Let $D$ be the uniform distribution over $[-1, 1]$ and $F$ be the bandlimit of the frequencies. We fix the family $\mathcal{F}$ in this section to be

$$\mathcal{F} = \left\{ f(x) = \sum_{j=1}^{k} v_j e^{2\pi \mathbf{i} \cdot f_j x} \middle| v_j \in \mathbb{C}, |f_j| \leq F \right\}.$$

The main result in this section is an estimation of the importance sampling of $x \in [-1, 1]$.

**Theorem 1.5.** *For any $x \in (-1, 1)$,*

$$\sup_{f \in \mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2} \lesssim \frac{k \log k}{1 - |x|}.$$

This directly improves $\kappa = \mathop{\mathbb{E}}_{x \in [-1,1]}[\sup_{f \in \mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2}]$ for signals with $k$-sparse Fourier transform, which is better than the condition number $\sup_{x \in [-1,1]} [\sup_{f \in \mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2}]$ used in [CKPS16].

**Theorem 9.1.** *For signals with k-sparse Fourier transform,*

$$\mathop{\mathbb{E}}_{x\in[-1,1]}\left[\sup_{f\in\mathcal{F}}\frac{|f(x)|^2}{\|f\|_D^2}\right]=O(k\log^2 k).$$

*Moreover, there exists a constant $c=\Theta(1)$ such that a distribution*

$$D_{\mathcal{F}}(x)=\begin{cases}\frac{c}{(1-|x|)\log k}, & \text{for } |x|\le 1-\frac{1}{k^3\log^2 k}\\ c\cdot k^3\log k, & \text{for } |x|>1-\frac{1}{k^3\log^2 k}\end{cases}$$

*guarantees for any* $f(x)=\sum_{j=1}^{k}v_j e^{2\pi \mathbf{i} f_j x}$ *and any* $x\in[-1,1]$, $\quad |f(x)|^2\cdot\dfrac{D(x)}{D_{\mathcal{F}}(x)}=O(k\log^2 k)\cdot\|f\|_D^2.$

We first state the condition number result in the previous work [CKPS16].

**Lemma 9.2** (Lemma 5.1 of [CKPS16])**.** *For any* $f(x)=\sum_{j=1}^{k}v_j e^{2\pi \mathbf{i} f_j x}$,

$$\sup_{x\in[-1,1]}\frac{|f(x)|^2}{\|f\|_D^2}=O(k^4\log^3 k).$$

We first show an interpolation lemma of $f(x)$ then finish the proof of Theorem 1.5.

**Claim 9.3.** *Given* $f(x)=\sum_{j=1}^{k}v_j e^{2\pi \boldsymbol{i} f_j\cdot x}$ *and* $\Delta$, *there exists* $l\in[2k]$ *such that for any* $t$,

$$|f(t+l\cdot\Delta)|^2\lesssim\sum_{j\in[2k]\setminus\{l\}}|f(t+j\cdot\Delta)|^2.$$

*Proof.* Given $k$ frequencies $f_1,\dots,f_k$ and $\Delta$, we set $z_1=e^{2\pi \boldsymbol{i} f_1\cdot\Delta},\dots,z_k=e^{2\pi\boldsymbol{i} f_k\cdot\Delta}$. Let $V$ be the linear subspace

$$\left\{\big(\alpha(0),\dots,\alpha(2k-1)\big)\in\mathbb{C}^{2k}\Big|\sum_{j=0}^{2k-1}\alpha(j)\cdot z_i^j=0,\forall i\in[k]\right\}.$$

Because the dimension of $V$ is $k$, let $\alpha_1,\dots,\alpha_k\in V$ be $k$ orthogonal coefficient vectors with unit length $\|\alpha_i\|_2=1$. From the definition of $\alpha_i$, we have

$$\begin{aligned}\sum_{j\in[2k]}\alpha_i(j)\cdot f(t+j\cdot\Delta)&=\sum_{j\in[2k]}\alpha_i(j)\sum_{j'\in[k]}v_{j'}\cdot e^{2\pi\boldsymbol{i} f_{j'}\cdot(t+j\Delta)}\\&=\sum_{j\in[2k]}\alpha_i(j)\sum_{j'\in[k]}v_{j'}\cdot e^{2\pi\boldsymbol{i} f_{j'}t}\cdot z_{j'}^j=\sum_{j'}v_{j'}\cdot e^{2\pi\boldsymbol{i} f_{j'}t}\sum_{j\in[2k]}\alpha_i(j)\cdot z_{j'}^j=0.\end{aligned}$$

Let $l$ be the coordinate in $[2k]$ with the largest weight $\sum_{i=1}^{k}|\alpha_i(l)|^2$. For every $i\in[k]$, from the above discussion,

$$-\alpha_i(l)\cdot f(t+l\cdot\Delta)=\sum_{j\in[2k]\setminus\{l\}}\alpha_i(j)\cdot f(t+j\cdot\Delta). \tag{15}$$

Let $A\in\mathbb{R}^{[k]\times[2k-1]}$ denote the matrix of the coefficients excluding the coordinate $l$, i.e.,

$$A=\begin{pmatrix}\alpha_1(0)&\cdots&\alpha_1(l-1)&\alpha_1(l+1)&\cdots&\alpha_1(2k-1)\\\alpha_2(0)&\cdots&\alpha_2(l-1)&\alpha_2(l+1)&\cdots&\alpha_2(2k-1)\\\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\\alpha_k(0)&\cdots&\alpha_k(l-1)&\alpha_k(l+1)&\cdots&\alpha_k(2k-1)\end{pmatrix}.$$

For the $k \times k$ matrix $A \cdot A^*$, its entry $(i, i')$ equals

$$\sum_{j \in [2k] \setminus \{l\}} \alpha_i(j) \cdot \overline{\alpha_{i'}(j)} = \langle \alpha_i, \alpha_{i'} \rangle - \alpha_i(l) \cdot \overline{\alpha_{i'}(l)} = 1_{i=i'} - \alpha_i(l) \cdot \overline{\alpha_{i'}(l)}.$$

Thus the eigenvalues of $A \cdot A^*$ are bounded by $1 + \sum_{i \in [k]} |\alpha_i(l)|^2$, which also bounds the eigenvalues of $A^* \cdot A$ by $1 + \sum_{i \in [k]} |\alpha_i(l)|^2$. From (15),

$$\sum_{i \in [k]} |\alpha_i(l) \cdot f(t + l \cdot \Delta)|^2 \le \lambda_{\max}(A^* \cdot A) \cdot \sum_{j \in [2k] \setminus \{l\}} |f(t + j \cdot \Delta)|^2$$
$$\Rightarrow \big( \sum_{i \in [k]} |\alpha_i(l)|^2 \big) \cdot |f(t + l \cdot \Delta)|^2 \le (1 + \sum_{i \in [k]} |\alpha_i(l)|^2) \cdot \sum_{j \in [2k] \setminus \{l\}} |f(t + j \cdot \Delta)|^2.$$

Because $l = \arg\max_{j \in [2k]} \big\{ \sum_{i \in [k]} |\alpha_i(j)|^2 \big\}$ and $\alpha_1, \ldots, \alpha_k$ are unit vectors, $\sum_{i \in [k]} |\alpha_i(l)|^2 \ge \sum_{i=1}^k \|\alpha_i\|_2^2 / 2k \ge 1/2$. Therefore

$$|f(t + l \cdot \Delta)|^2 \le 3 \sum_{j \in [2k] \setminus \{l\}} |f(t + j \cdot \Delta)|^2.$$

□

**Corollary 9.4.** *Given $f(x) = \sum_{j=1}^k v_j e^{2\pi \boldsymbol{i} f_j \cdot x}$, for any $\Delta$ and $t$,*

$$|f(t)|^2 \lesssim \sum_{i=1}^{2k} |f(t + i\Delta)|^2 + \sum_{i=1}^{2k} |f(t - i\Delta)|^2.$$

Next we finish the proof of Theorem 1.5.

*Proof of Theorem 1.5.* We assume $t = 1 - \epsilon$ for an $\epsilon \le 1$ and integrate $\Delta$ from 0 to $\epsilon/2k$:

$$\begin{aligned}
\varepsilon/2k \cdot |f(t)|^2 &\lesssim \int_{\Delta=0}^{\varepsilon/2k} \sum_{i=1}^{2k} |f(t + i\Delta)|^2 + \sum_{i=1}^{2k} |f(t - i\Delta)|^2 \mathrm{d}\Delta \\
&= \sum_{i \in [1, \ldots, 2k]} \int_{\Delta=0}^{\varepsilon/2k} |f(t + i\Delta)|^2 + |f(t - i\Delta)|^2 \mathrm{d}\Delta \\
&\lesssim \sum_{i \in [1, \ldots, 2k]} \frac{1}{i} \cdot \int_{\Delta'=0}^{\varepsilon \cdot i/2k} |f(t + \Delta')|^2 \mathrm{d}\Delta' + \sum_{i \in [1, \ldots, 2k]} \frac{1}{i} \cdot \int_{\Delta'=0}^{\varepsilon \cdot i/2k} |f(t - \Delta')|^2 \mathrm{d}\Delta' \\
&\lesssim \sum_{i \in [1, \ldots, 2k]} \frac{1}{i} \cdot \int_{\Delta'=-\varepsilon}^{\varepsilon} |f(t + \Delta')|^2 \mathrm{d}\Delta' \\
&\lesssim \log k \cdot \int_{x=-1}^{1} |f(x)|^2 \mathrm{d}x.
\end{aligned}$$

From all discussion above, we have $|f(1 - \varepsilon)|^2 \lesssim \frac{k \log k}{\varepsilon} \cdot \mathop{\mathbb{E}}_{x \in [-1,1]} [|f(x)|^2]$. □

*Proof of Theorem 9.1.* We bound

$$\begin{aligned}
\kappa &= \mathop{\mathbb{E}}_{x\in[-1,1]}[\sup_{f\in\mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2}] \\
&= \frac{1}{2}\int_{x=-1}^{1} \sup_{f\in\mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2} \mathrm{d}x \\
&\lesssim \int_{x=-1+\varepsilon}^{1-\varepsilon} \sup_{f\in\mathcal{F}} \frac{|f(x)|^2}{\|f\|_D^2} \mathrm{d}x + \varepsilon\cdot k^4\log^3 k && \text{from Lemma 9.2} \\
&\lesssim \int_{x=-1+\varepsilon}^{1-\varepsilon} \frac{k\log k}{1-|x|}\mathrm{d}x + \varepsilon\cdot k^4\log^3 k && \text{from Theorem 1.5} \\
&\lesssim k\log k\cdot \log\frac{1}{\varepsilon} + \varepsilon\cdot k^4\log^3 k \lesssim k\log^2 k
\end{aligned}$$

by choosing $\varepsilon = \frac{1}{k^3\log k}$. Next we define $D_{\mathcal{F}}(x) = D(x)\cdot \frac{\sup_{f\in\mathcal{F},f\neq 0}\frac{|f(x)|^2}{\|f\|_D^2}}{\kappa}$. The description of $D_{\mathcal{F}}(x)$ follows the upper bound of $\sup_{f\in\mathcal{F},f\neq 0}\frac{|f(x)|^2}{\|f\|_D^2}$ in Lemma 9.2 and Theorem 1.5. From Claim 6.4, its condition number is $\kappa = O(k\log^2 k)$. □

Before we show a sample-efficient algorithm, we state the following version of the Chernoff bound that will used in this proof.

**Lemma 9.5** (Chernoff Bound [Che52, Tar09] )**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables. Assume that $0 \le X_i \le 1$ always, for each $i \in [n]$. Let $X = X_1 + X_2 + \cdots + X_n$ and $\mu = \mathbb{E}[X] = \sum_{i=1}^{n}\mathbb{E}[X_i]$. Then for any $\varepsilon > 0$,*

$$\Pr[X \ge (1+\varepsilon)\mu] \le \exp(-\frac{\varepsilon^2}{2+\varepsilon}\mu) \text{ and } \Pr[X \ge (1-\varepsilon)\mu] \le \exp(-\frac{\varepsilon^2}{2}\mu).$$

**Corollary 9.6.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables in $[0, R]$ with expectation 1. For any $\varepsilon < 1/2$, $X = \frac{\sum_{i=1}^n X_i}{n}$ with expectation 1 satisfies*

$$\Pr[|X-1| \ge \varepsilon] \le 2\exp(-\frac{\varepsilon^2}{3}\cdot\frac{n}{R}).$$

Finally, we provide a relatively sample-efficient algorithm to recover $k$-Fourier-sparse signals. Applying the same proof with uniform samples would require a $K/\kappa = O(k^3)$ factor more samples.

**Corollary 9.7.** *For any $F > 0, T > 0, \varepsilon > 0$, and observation $y(x) = \sum_{j=1}^{k} v_j e^{2\pi\mathbf{i}f_j x} + g(x)$ with $|f_j| \le F$ for each $j$, there exists a non-adaptive algorithm that takes $m = O(k^4\log^3 k + k^2\log^2 k\cdot \log\frac{FT}{\varepsilon})$ random samples $t_1, \ldots, t_m$ from $D_{\mathcal{F}}$ and outputs $\widetilde{f}(x) = \sum_{j=1}^{k}\widetilde{v}_j e^{2\pi\mathbf{i}\widetilde{f}_j x}$ satisfying*

$$\mathop{\mathbb{E}}_{x\sim[-T,T]}\left[|\widetilde{f}(x)-f(x)|^2\right] \lesssim \mathop{\mathbb{E}}_{x\sim[-T,T]}\left[|g(x)|^2\right] + \varepsilon \mathop{\mathbb{E}}_{x\sim[-T,T]}\left[|f(x)|^2\right] \text{ with probability 0.9.}$$

*Proof.* We first state the main tool from the previous work. From Lemma 2.1 in [CKPS16], let $N_f = \frac{\varepsilon}{T\cdot k^{Ck^2}}\cdot\mathbb{Z}\cap[-F,F]$ denote a net of frequencies for a constant $C$. For any signal $f(x) =$

**Algorithm 5** Recover $k$-sparse FT

**procedure** SPARSEFT$(y, F, T, \varepsilon)$
    $m \leftarrow O(k^4 \log^3 k + k^2 \log^2 k \log \frac{FT}{\varepsilon})$
    Sample $t_1, \ldots, t_m$ from $D_{\mathcal{F}}$ independently
    Set the corresponding weights $(w_1, \ldots, w_m)$ and $S = (t_1, \ldots, t_m)$
    Query $y(t_1), \ldots, y(t_m)$ from the observation $y$
    $N_f \leftarrow \frac{\varepsilon}{T \cdot k^{Ck^2}} \cdot \mathbb{Z} \cap [-F, F]$ for a constant $C$
    **for** all possible $k$ frequencies $f'_1, \ldots, f'_k$ in $N_f$ **do**
        Find $h(x)$ in $\mathrm{span}\{e^{2\pi \mathbf{i} \cdot f'_1 x}, \ldots, e^{2\pi \mathbf{i} \cdot f'_k x}\}$ minimizing $\|h - y\|_{S,w}$
        Update $\widetilde{f} = h$ if $\|h - y\|_{S,w} \leq \|\widetilde{f} - y\|_{S,w}$
    **end for**
    Return $\widetilde{f}$.
**end procedure**

$\sum_{j=1}^{k} v_j e^{2\pi \mathbf{i} f_j(x)}$, there exists a $k$-sparse signal

$$f'(x) = \sum_{j=1}^{k} v'_j e^{2\pi \mathbf{i} f'_j(x)} \text{ satisfying } \|f - f'\|_D \leq \varepsilon \|f\|_D$$

whose frequencies $f'_1, \ldots, f'_k$ are in $N_f$. We rewrite $y = f + g = f' + g'$ where $g' = g + f - f'$ with $\|g'\|_D \leq \|g\|_D + \varepsilon \|f\|_D$. Our goal is to recover $f'$.

We construct a $\delta$-net with $\delta = 0.05$ for

$$\left\{ h(x) = \sum_{j=1}^{2k} v_j e^{2\pi \mathbf{i} \cdot \widehat{h}_j x} \middle| \|h\|_D = 1, \widehat{h}_j \in N_f \right\}.$$

We first pick $2k$ frequencies $\widehat{h}_1, \ldots, \widehat{h}_{2k}$ in $N_f$ then construct a $\delta$-net on the linear subspace $\mathrm{span}\{e^{2\pi \mathbf{i} \widehat{h}_1 x}, \ldots, e^{2\pi \mathbf{i} \widehat{h}_{2k} x}\}$. Hence the size of our $\delta$ net is

$$\binom{\frac{4FT \cdot k^{Ck^2}}{\varepsilon}}{2k} \cdot (12/\delta)^{2k} \leq (\frac{4FT \cdot k^{Ck^2}}{\varepsilon \cdot \delta})^{3k}.$$

Now we consider the number of random samples from $D_{\mathcal{F}}$ to estimate signals in the $\delta$-net. Based on the condition number of $D_{\mathcal{F}}$ in Theorem 9.1 and the Chernoff bound of Corollary 9.6, a union bound over the $\delta$-net indicates

$$m = O(\frac{k \log^2 k}{\delta^2} \cdot \log |\mathrm{net}|) = O\left( \frac{k \log^2 k}{\delta^2} \cdot (k^3 \log k + k \log \frac{FT}{\varepsilon \delta}) \right)$$

random samples from $D_{\mathcal{F}}$ would guarantee that for any signal $h$ in the net, $\|h\|_{S,w}^2 = (1 \pm \delta) \|h\|_D^2$. From the property of the net,

$$\text{for any } h(x) = \sum_{j=1}^{2k} v_j e^{2\pi \mathbf{i} h_j(x)} \text{ with } \widehat{h}_j \in N_f, \quad \|h\|_{S,w}^2 = (1 \pm 2\delta) \|h\|_D^2,$$

which is sufficient to recover $f'$.

We present the algorithm in Algorithm 5 and bound $\|f-\widetilde{f}\|_D$ as follows. The expectation of $\|f-\widetilde{f}\|_D$ over the random samples $S=(t_1,\ldots,t_m)$ is

$$\begin{aligned}\|f-f'\|_D+\|f'-\widetilde{f}\|_D &\le \|f-f'\|_D+1.1\|f'-\widetilde{f}\|_{S,w}\\ &\le \|f-f'\|_D+1.1(\|f'-y\|_{S,w}+\|y-\widetilde{f}\|_{S,w})\\ &\le \|f-f'\|_D+1.1(\|g'\|_{S,w}+\|y-f'\|_{S,w})\\ &\le \varepsilon\|f\|_D+2.2(\|g\|_D+\varepsilon\|f\|_D).\end{aligned}$$

From the Markov inequality, with probability 0.9, $\|f-\widetilde{f}\|_D \lesssim \varepsilon\|f\|_D+\|g\|_D$. $\square$


## Acknowledgements

The authors would like to thank Adam Klivans and David Zuckerman for many helpful comments about this work.


## References


[AKM+17] Haim Avron, Michael Kapralov, Cameron Musco, Christopher Musco, Ameya Velingker, and Amir Zandieh. Random fourier features for kernel ridge regression: Approximation bounds and statistical guarantees. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, pages 253–262, 2017.

[AZLSW17] Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal discrete optimization for experimental design: A regret minimization approach. *arXiv preprint arXiv:1711.05174*, 2017.

[BDMI13] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal coresets for least-squares regression. *IEEE transactions on information theory*, 59(10):6880–6892, 2013.

[BM86] Y. Bresler and A. Macovski. Exact maximum likelihood parameter estimation of superimposed exponential signals in noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5):1081–1089, Oct 1986.

[BSS12] Joshua Batson, Daniel A Spielman, and Nikhil Srivastava. Twice-ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012.

[CDL13] Albert Cohen, Mark A Davenport, and Dany Leviatan. On the stability and accuracy of least squares approximations. *Foundations of computational mathematics*, 13(5):819–834, 2013.

[Che52] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23:493–507, 1952.

[CKNS15] Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, pages 1090–1098, 2015.

[CKPS16] Xue Chen, Daniel M. Kane, Eric Price, and Zhao Song. Fourier-sparse interpolation without a frequency gap. In *FOCS 2016*, 2016.

[DMM08] Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

[DW17] Michal Derezinski and Manfred K Warmuth. Unbiased estimates for linear regression via volume sampling. In *Advances in Neural Information Processing Systems*, pages 3087–3096, 2017.

[DWH18] Michal Derezinski, Manfred K Warmuth, and Daniel Hsu. Tail bounds for volume sampled linear regression. *arXiv preprint arXiv:1802.06749*, 2018.

[Fan61] Robert Fano. *Transmission of information; a statistical theory of communications.* Cambridge, Massachusetts, M.I.T. Press, 1961.

[Har28] Ralph Hartley. Transmission of information. *Bell System Technical Journal*, 1928.

[HS16] Daniel Hsu and Sivan Sabato. Loss minimization and parameter estimation with heavy tails. *The Journal of Machine Learning Research*, 17(1):543–582, 2016.

[LS15] Yin Tat Lee and He Sun. Constructing linear-sized spectral sparsification in almost-linear time. In *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, FOCS '15, pages 250–269. IEEE Computer Society, 2015.

[Mah11] Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

[MI10] Malik Magdon-Ismail. Row sampling for matrix algorithms via a non-commutative bernstein bound. *arXiv preprint arXiv:1008.0587*, 2010.

[Moi15] Ankur Moitra. The threshold for super-resolution via extremal functions. In *STOC*, 2015.

[MV74] H.L. Montgomery and R.C. Vaughan. Hilbert's inequality. *Journal of the London Mathematical Society*, s2-8(1):73–82, 1974.

[RW12] Holger Rauhut and Rachel Ward. Sparse legendre expansions via $\ell_1$-minimization. *Journal of approximation theory*, 164(5):517–533, 2012.

[Sha49] Claude Shannon. Communication in the presence of noise. *Proc. Institute of Radio Engineers*, 37(1):10–21, 1949.

[SM14] Sivan Sabato and Remi Munos. Active regression by stratification. In *Advances in Neural Information Processing Systems*, pages 469–477, 2014.

[SWZ19] Zhao Song, David P Woodruff, and Peilin Zhong. Relative error tensor low rank approximation. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2772–2789. SIAM, 2019.

[Tar09] Robert E. Tarjan. Lecture 10: More chernoff bounds, sampling, and the chernoff + union bound. *Princeton Class Notes, Probability and Computing*, pages 1–9, 2009.

[Tro12] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012.

[War15] Rachel Ward. Importance sampling in signal processing applications. In *Excursions in Harmonic Analysis, Volume 4*, pages 205–228. Springer, 2015.

[Woo14] David P Woodruff. Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*, 2014.