Hard-Margin Active Linear Regression

Elad Hazan EHAZAN@IE.TECHNION.AC.IL
Technion, Haifa, Israel

Zohar Karnin ZKARNIN@YAHOO.COM
Yahoo Labs, Haifa, Israel

Abstract

We consider the fundamental problem of linear regression in which the designer can actively choose observations. This model naturally captures various experiment design settings in medical experiments, ad placement problems and more. Whereas previous literature addresses the soft-margin or mean-square-error variants of the problem, we consider a natural machine learning hard-margin criterion. In this setting, we show that active learning admits significantly better sample complexity bounds than the passive learning counterpart, and give efficient algorithms that attain near-optimal bounds.

1. Introduction

In this paper we consider a problem of experiment design from an active learning viewpoint. The setting at hand is a linear regression problem where the error is not measured in the standard mean square loss function, but rather in the natural 0-1 loss setting. That is, a data point will induce a loss of cost one if the regressor produces an error larger than some threshold and zero otherwise.

The 0-1 loss scenario is motivated by numerous real-life scenarios. The first example is medical experiments: consider a scientist attempting to predict a condition according to patient attributes. In this setting, different patients have different attributes, or features, and the experiment designer goal is to test a minimal number of patients before being able to successfully predict the condition of future patients based on their attributes. In a regression setting it is likely that predicting the outcome up to some small error will induce no loss, while erring by slightly more or much more will have the same affect of mistreating the patient.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

Another example is a recommendation or ad serving system where the objective is to serve relevant items to users. These items could be movies, articles, webpages, advertisements, etc.. A very common approach to the problem is to assume that the relevance of an item for a user is determined as a bi-linear function of the features of the user and item. Consider the case where the users are not given by their identity but only as a list of attributes such as gender, location, age, etc (a common scenario in web related applications). By considering categories of the items the problem is reduced to learning multiple regressors. The true objective of these regressors is to present relevant items to the users. The underlying objective function aligns perfectly with the 0-1 loss setting; an item is either relevant or not relevant, as indicated by the user when choosing whether to view / click / purchase the item.

Formally we can model this experiment design setting as follows: a data set (of patients/users/etc) is associated with a set of feature, or attribute, vectors $\mathcal{K} \subseteq \mathbb{R}^d$. An experiment with patient $x \in \mathcal{K}$ corresponds to a noisy measurement, or "query", $\langle x, w^* \rangle + \varepsilon_x$, where w^* is the optimal underlying linear model and ε_x is a zero mean and bounded variance noise. The goal of the experiment designer is to adaptively choose and measure a sequence of experiments with the goal of attaining, with high probability, an ε-close estimate of entire data set [1]

The difficulty presented in this problem is two-fold: first there is the statistical-geometrical difficulty of exploration - which experiments should be conducted in order to accurately predict the values over all data points? Second - there is the optimization difficult of efficient computation of the optimal linear regressor in an adaptive setting. These two questions are obviously related - the ideal method should interpolate between measurements, optimization to identify the areas that require more exploration, more experiments

[1] By standard techniques this can be generalised to a distributional setting, in which patients x arrive from an unknown and i.i.d distribution $x \sim \mathcal{D}$ and the requirement is that with probability at least $1 - \delta$ the regressor errs by less than ε.

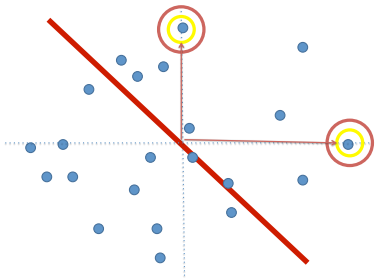

Figure 1. Illustration of a discrete experiment design: the main problem is to pick the most informative experiments. In this case there is an optimal orthonormal basis that spans the space - the two highlighted points.

et cetera.

Clearly, if the data lives on the real line, $\Omega(\frac{1}{\varepsilon^2})$ noisy measurements are required to obtain an ε-approximate regressor. This is generalized to $\Omega(\frac{d}{\varepsilon^2})$ experiments for data that lives in d-dimensional Euclidean space. The question becomes interesting once we have very large data sets. Suppose there are $n \gg d$ patients for which predictions are needed, **can we regress on the entire data making less than $\Theta(n)$ experiments?**

Clearly in a passive-learning setting the answer is no. The distribution over the data could be so skewed as to give no or very little information about an important part of space. However - an active learner can potentially exploit the geometric structure of the data and attain much better bounds.

Our main contribution is indeed an active/adaptive algorithm for experiment design in this framework that achieves the following guarantee: after querying $\tilde{O}\left(\frac{d}{\varepsilon^2}\right)$ carefully-chosen experiments, the algorithm returns a vector that is ε-accurate for all data points. Furthermore, our algorithm is computationally efficient and can be implemented to run in near-linear running time given a low-variance exploration basis for the experiment space (that can be constructed in polynomial time).

1.1. Related work

The natural problem we address has been considered before in many variants in two main communities: experiment design and active learning.

Experiment design: In the statistical field called *optimal design of experiments*, or just *optimal design* (Atkinson & Donev, 1992; Wu, 1978) , a statistician is faced with the task of choosing experiments to perform from a given pool, with the goal of producing the optimal result within the budget constraint.

Formally, consider a pool of possible experiments denoted $x_1, ..., x_n \in \mathbb{R}^d$. The goal of the designer is to choose a distribution over the pool of experiments, such that experiments chosen according to this distribution produce a hypothesis $\hat{\mathbf{w}}$ that is as close as possible to the true linear function behind the data. The distance between the hypothesis and true linear function can be measured in different ways, each corresponding to a different *optimality criteria*. The common property of the criteria is that they all minimize the variance of the hypothesis. Since the variance is not a scalar but a $d \times d$ matrix, the different criteria differ by the fact that each one minimizes a different function $\Phi : \mathbb{R}^{d \times d} \to \mathbb{R}$ over the covariance matrix. Common criteria are the A-, D-, and E-optimality criteria. D-optimality, minimizes the determinant of the covariance matrix, and thus minimizes the volume of the confidence region. In A-optimality the trace of the covariance matrix, i.e. the total variance of the parameter estimates, is minimized. E-optimality minimizes the maximum eigenvalue of the covariance matrix, and thus minimizes the size of the major axis of the confidence region.

The above criteria do not directly characterize the quality of predictions on test data. A common criterion that directly takes the test data into account is that of G-optimality. Here the goal is to minimize the maximum variance of the predicted values. In other words, by denoting $\mathrm{Var}_S(x_i)$ the variance of the prediction of x_i after querying the points of S, the goal in G-optimality is to minimize $\max_i \mathrm{Var}_S(x_i)$. G-optimality and D-optimality are closely related in the sense that an exact solution to one is the solution to the other (see e.g. (Spruill & Studden, 1979)). Note that this criterion is similar to our objective. The difference is twofold. First, we do not aim to minimize the variance but to obtain a high probability bound. Second, in optimal design the quality of the distribution is measured when the budget tends to infinity. Specifically, notice that for a distribution over the possible experiments, rather than a deterministic subset of them, the corresponding covariance matrix is random. The discussed minimizations are done over the expected covariance matrix, where the expectation is taken over the subset of chosen experiments. When the budget tends to infinity the actual covariance matrix is close w.h.p to its expected counterpart. We call this the *infinite budget setting*. Ours is a finite budget setting where one does not aim to provide a distribution over the possible experiments but a deterministic subset of them of a fixed size.

For the finite budget setting various relaxations have been considered in the statistical literature, usually without an approximation guarantee. Our method differs from previous works of this spirit by: First, we do not impose a hard-budget constraint of experiments, but rather bound the number of experiments as a function of the desired approximation guarantee. Second, we obtain a computationally efficient algorithm with provable optimality results. Finally,

as an added bonus our solution has the property of choosing very few data points to explore, potentially much less than the budget. A motivating example for this property is the medical experiment design. Here a data point is a human subject and it is more realistic to have few volunteers being thoroughly tested on as opposed to performing few tests over many volunteers.

Our setting is arguably more natural for the medical-patient-experiment motivating example; in general there are numerous examples where the budget of experiments is not fixed but rather the tolerable error. Of equal importance is the fact that our setting allows to derive efficient algorithms with rigorous theoretical guarantees.

A related and recently popular model is called random design (Hsu et al., 2012; Audibert & Catoni, 2010; Györfi, 2002; Audibert & Catoni, 2011). In this setting the designer is given a set of measurements $\{x_i, y_i | i \in [n]\}$ for $x_i \in \mathbb{R}^d$ drawn from an unknown distribution $\mathcal{D}$. The goal is to predict as well as the best linear predictor measured according to the mean square error, i.e. minimize

$$\mathbf{E}_{(x,y)\in\mathcal{D}}\left[(x^\top w - y)^2 - (x^\top w^* - y)^2\right]$$

where w^* is the optimal linear regressor. Various other performance metrics have been considered in the referenced papers, i.e. measuring the norm of the regressor vs. the optimal regressor in a norm proportional to the covariance matrix. However, in this setting an expected error is the criterion vs. our criterion of worst-case, or a high confidence bound on the error[2], which is more suitable for some experiment design settings.

Active learning: The most well-studied setting in active learning is pool-based active learning (McCallum & Nigam, 1998), in which the learner has access to a pool of examples, and can iteratively query labels of particular examples of her choice. Compared to passive learning, in which labelled examples are drawn from a fixed unknown distribution, known active learning algorithms can attain a certain generalization error guarantee albeit observing exponentially fewer labelled examples, e.g. (Cohn et al., 1994; Dasgupta et al., 2009; Hanneke, 2007; Balcan et al., 2009), under certain assumptions such as special hypothesis classes, realizability or large-margin. Active learning with noise is a much less studied topic: (Balcan et al., 2009) give an exponential improvement over passive learning of linear threshold functions, but under the condition that the noise is smaller than the desired accuracy. Real valued active learning with a soft-margin criteria was addressed in (Ganti & Gray, 2012). The reader is referred to (Dasgupta & Langford, 2009) for a more detailed survey of active learning literature.

[2]Our results though stated as a worst case error can be generalized to a high probability solution in the random design scenario

2. Preliminaries

Let $\mathbb{R}^d$ be the space of d dimensional vectors over the reals. Throughout the paper we denote the set of n data points as $\mathcal{K} = \{x_1, \ldots, x_n\}$. We denote by $w^* \in \mathbb{R}^d$ the hidden regressor. A *query* at point x, reveals a noisy output denoted by $\widehat{\langle x, w^* \rangle}$ consisting of the true inner product $\langle x, w^* \rangle$ plus a zero mean noise. We operate under the following assumptions: First, the different noise elements of different queries are independent. Second, the output of $\widehat{\langle x, w^* \rangle}$ is always bounded in $[-1, 1]$; furthermore, all of the data points $x \in \mathcal{K}$ and the hidden vector w^* are in the Euclidean unit sphere. Relaxations of these assumptions are in many cases possible via standard techniques, and in this extended abstract we focus on the core problem. The objective in the *Hard-Margin Active Linear Regression* problem, or ALR in short, is to to learn a regressor $w \in \mathbb{R}^d$ where it holds with probability at least $1 - \delta$ that

$$\forall x \in \mathcal{K}, \;\; |\langle x, w \rangle - \langle x, w^* \rangle| < \varepsilon \tag{1}$$

An algorithm for this problem has two measurements. First and more important for our setting is the query complexity. That is, how many queries did the algorithm use? The second measure is the computational complexity which is the running time.

Before diving into our algorithmic results, we mention here a lower bound on the query complexity of ALR that can be derived by reduction to known lower bounds for the stochastic multi-armed bandit problem. We prove the following theorem in section 3:

Theorem 1 (Query Complexity Lower Bound)**.** *Any algorithm for the ALR problem requires $\Omega(d \log(1/\delta)/\varepsilon^2)$ measurements in order to compute a linear regressor with error of at most ε with probability larger than $1 - \delta$.*

2.1. Volumetric Spanners

A non-active solution to the ALR problem would simply experiment over all data points sufficiently many times, and apply any optimization method for finding the optimal regressor over the measurements. This is of course unsatisfactory - an active learner would want to exploit similarities in data and closeness in feature space to achieve faster learning.

An important technical tool for exploiting such structure is an informative exploration basis. That is, a subset of points from $\mathcal{K}$ such that querying the points of this set will provide the most information for all of the points in $\mathcal{K}$. To this end we use a recently devised geometric object called a *volumetric spanner* (Hazan et al., 2013). A volumetric spanner is a finite subset of $\mathcal{K}$ such that any point in $\mathcal{K}$ is spanned by it with small coefficients, or formally:

Definition 1. *Let $\mathcal{K} \subseteq \mathbb{R}^d$ and let $S = \{v_1, \ldots, v_{|S|}\}$ be*

a (multi-)subset[3] *of $\mathcal{K}$. Let V be the $d \times |S|$ matrix whose columns are the vectors of S. For a vector x let*

$$\|x\|_{\mathcal{E}(S)} = \sqrt{x(VV^\top)^{-1}x}$$

The set S is a volumetric spanner *of $\mathcal{K}$ when for all $x \in \mathcal{K}$ it holds that $\|x\|_{\mathcal{E}(S)} \leq 1$*

We later show that the properties of the volumetric spanner are exactly those that are required of an exploration set. That is, when querying the points of a set $S \subseteq \mathcal{K}$, the quality of the corresponding regressor w.r.t a data point x is exactly proportional to $\|x\|_{\mathcal{E}(S)}$. The query complexity of the algorithm will be determined by the size of the set S. In (Hazan et al., 2013) it is proven that such spanners of cardinality $O(d)$ exist for any compact set $\mathcal{K}$. Furthermore, for finite sets $\mathcal{K} \subseteq \mathbb{R}^d$ of cardinality n, such spanners can be found in polynomial time in n and d [4]. Formally:

Theorem 2. *[(Hazan et al., 2013)] Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a finite set of size n. There exist an algorithm that constructs a volumetric spanner for $\mathcal{K}$ of cardinality at most $12d$ whose running time is $O(n^{3.5} + n^3 d + d^5)$. An alternative algorithm for the problem exits with running time of $O(nd^2)$ achieving a volumetric spanner of size $O(d\log(d)\log(n))$.*

3. A lower bound for passive linear regression

In this section we provide an example for a set X where the passive learning algorithm must use $\Omega(\frac{n}{\varepsilon^2})$ observations to obtain a regressor w with additive error of at most ε on all of the data points. We start by better defining the passive setup. Here, a query returns a random point x chosen uniformly from the set $\mathcal{K}$ and a noisy measurement of $\langle w^*, x\rangle$. As before we assume that all points, including w^* are contained in the ℓ_2 unit sphere.

Our set $\mathcal{K}$ is defined in the following manner. Let $Y \subseteq \mathbb{R}^d$ be an arbitrary set of size $n-1$ such that for all $x \in Y$, $\langle x, e_1\rangle = 0$. Let $\mathcal{K} = Y \cup \{e_1\}$.

Theorem 3. *Any algorithm in the passive setting achieving an additive error of at most ε in all of the data points of $\mathcal{K}$ whose success probability is $1-\delta$ requires $\Omega(\log(1/\delta)n/\varepsilon^2)$ queries.*

The theorem is an immediate corollary of the following lemma.

Lemma 4. *For $\mathcal{K}$ defined above, any policy distinguishing between the case where $\langle w^*, e_1\rangle = -\varepsilon$ and $\langle w^*, e_1\rangle = \varepsilon$ with probability larger than $1-\delta$ must use $\Omega(n\log(1/\delta)/\varepsilon^2)$ queries.*

[3] S can be a multi-set, meaning it can contain several copies of the same element

[4] Extensions to infinite sets is also addressed in (Hazan et al., 2013), but outside the scope of this paper

Proof. We begin by mentioning that a query of a point x where $\langle x, e_1\rangle = 0$ provides no information to the sign of $\langle w^*, e_1\rangle$, hence does not help distinguish between the two hypotheses. The following lemma provides a lower bound to the number of queries at point e_1 required to estimate the $\langle w^*, e_1\rangle$ up to a sufficiently small additive error and with sufficient confidence. It is a folklore lemma in statistics and appears e.g., in (Mannor & Tsitsiklis, 2004) in a much more general form.

Lemma 5 (Theorem 1 of (Mannor & Tsitsiklis, 2004))**.** *Let $\mathcal{D}$ be a distribution over $[-1,1]$. Let $\varepsilon > 0$ be such that for $X \sim \mathcal{D}$, $|\mathbf{E}[X]| \geq \varepsilon$. Let T be the expected number of queries required by any algorithm that queries i.i.d copies of $X \sim \mathcal{D}$ until being able to distinguish, with probability at least $1-\delta$ between the cases $\mathbf{E}[X] \leq -\varepsilon$ and $\mathbf{E}[X] \geq \varepsilon$. Then for universal constants $\varepsilon_0 > 0$, $\delta_0 > 0$, c_1, c_2 it holds that if $\varepsilon < \varepsilon_0$ and $\delta < \delta_0$ then $T \geq \frac{c_1 \log(c_2/\delta)}{\varepsilon^2}$.*

It follows that the expected number of queries needed in order to distinguish between the two hypotheses with probability $\geq 1-\delta$ is at least $\frac{c_1 n \log(c_2/\delta)}{\varepsilon^2}$, as the probability of observing a query to the inner product with e_1 is $1/n$. □

4. Our ALR solution

In this section we provide a high level description of our solution. In the following sections we prove the following.

Theorem 6. *There exists a solution to the ALR problem with success probability of at least $1-\delta$ with the following properties: (1) The solution requires a preprocessing stage of $O(n^{3.5} + dn^3 + d^5)$ (2) it's running time (after preprocessing) is $\tilde{O}\left(\frac{nd\log(1/\delta)}{\varepsilon^2}\right)$ and (3) it's query complexity is at most $O\left(\frac{d\log(n)\log(1/\delta)}{\varepsilon^2}\right)$. An alternative algorithm exists with a preprocessing time of $O(nd^2)$ requiring an additional factor of $\log(n)\log(d)$ for the number of queries.*

The intuition behind the algorithm is the following. We begin with a preprocessing stage of computing a volumetric spanner S for the set of points $\mathcal{K}$. Given this spanner we can implement a procedure that outputs, for all of the points of $\mathcal{K}$ simultaneously, an unbiased estimator of $\langle w^*, x\rangle$ with variance of at most $|S|$. To demonstrate the usefulness of this estimator, consider averaging $|S|\log(n/\delta)/\varepsilon^2$ i.i.d outputs of S. Standard concentration bound show that w.p at least $1-\delta$ the estimates of all points in $\mathcal{K}$ are correct up to an additive error of ε. Rather than computing a noisy output for w^* on the points and recovering a hypothesis w from that we use a technique by (Clarkson et al., 2012) that given an oracle for a function over a set of data points constructs a hypothesis w using a small number of queries to the oracle.

4.1. Constructing a low variance estimator

In the following section we provide an algorithm that requires a black box providing a noisy estimate of $\langle x, w^* \rangle$ to all of the data point of $\mathcal{K}$ simultaneously. The intuition behind the algorithm is that given sufficiently many queries to the noisy estimator, a union bound argument can ensure an accurate estimate in all of the data points simultaneously. In this section we begin with the description of this black box providing the estimates.

The main tool used for this 'all-point-estimator' is the volumetric spanner. Algorithm 1 provides the formal description of the method used to obtain the estimates.

Algorithm 1 **Sample**$(\mathcal{K})$

1: Input: set $\mathcal{K} = \{x_1, ..., x_n\}$, Volumetric spanner for $\mathcal{K}$ denoted S, and measurement oracle that given x returns an unbiased estimator $\widehat{\langle x, w^* \rangle}$ with variance at most one for some fixed w^*.
2: Choose a point $v_j \in S$ uniformly at random, query its inner product $\hat{\ell} = \widehat{\langle v_j, w^* \rangle}$
3: let V be the $d \times |S|$ matrix whose columns are the elements of S, and let $V^\dagger \in \mathbb{R}^{|S| \times d}$ be its Moore-Penrose pseudo inverse.
4: For $x \in \mathcal{K}$, let $\alpha_x = V^\dagger x$ and let $\hat{\ell}_x \leftarrow (\alpha_x)_j \hat{\ell} \cdot |S|$
5: **return** estimates $\{\hat{\ell}_x\}$

The following lemma provides the analysis of Algorithm 1.

Lemma 7. *Algorithm 1 queries a single point from $\mathcal{K}$. Its estimates have the properties of (1) being unbiased and (2) have a variance of at most $12d$. More formally, we have*

$$\forall x \in \mathcal{K} \,.\, \mathbf{E}[\hat{\ell}_x] = \langle x, w^* \rangle \;,\; \mathbf{Var}(\hat{\ell}_x) \leq |S|$$

Proof.

$$\begin{aligned} \mathbf{E}\left[\hat{\ell}_x\right] &= \textstyle\sum_{j \in S} \Pr[v_j] \cdot (\alpha_x)_j \mathbf{E}\widehat{\langle v_j, w^* \rangle} \cdot |S| \\ &= \textstyle\sum_{j \in S} (\alpha_x)_j \mathbf{E}\widehat{\langle v_j, w^* \rangle} \\ &= (V^\dagger x)^T V^T w^* = \langle x, w^* \rangle \end{aligned}$$

For the variance, recall that $x \in \mathcal{K}$ and S is a volumetric spanner of $\mathcal{K}$ indicates that $\|\alpha_x\|_2 \leq 1$:

$$\begin{aligned} \mathbf{E}[\hat{\ell}_x^2] &= \textstyle\sum_{j \in S} \Pr[v_j] \cdot (\alpha_x)_j^2 \mathbf{E}\left[\widehat{\langle v_j, w^* \rangle}^2\right] \cdot |S|^2 \\ &\leq |S| \textstyle\sum_{j \in S} (\alpha_x)_j^2 \leq |S| \end{aligned}$$

By Theorem 2 we can efficiently construct volumetric spanners of size $|S| = 12d$. □

4.2. Algorithm and its analysis

In this section we present an algorithm for the ALR problem, following the primal-dual paradigm of (Clarkson et al., 2012). It assumes an oracle to a procedure **Sample**$(\mathcal{K})$ that returns a vector of length $|\mathcal{K}|$ whose entries are unbiased estimators of $\langle x, w^* \rangle$, for all $x \in \mathcal{K}$ whose variance is upper bounded by $\tilde{O}(d)$. Recall that such a procedure was given in Section 4.1.

To avoid extraneous notions we will assume henceforth w.l.o.g that $\mathcal{K}$ is symmetric meaning that $x \in \mathcal{K}$ iff $-x \in \mathcal{K}$. This is without loss of generality since an unbiased estimator for $\langle -x, w^* \rangle$ is obtained by negating the estimator for $\langle x, w^* \rangle$.

Consider the following mathematical program:

$$\min_{\|w\| \leq 1} g(w) \;\; s.t. \qquad \begin{aligned} g(w) &= \max_{x \in \mathcal{K}} c_x(w) \\ c_x(w) &= \langle x, w \rangle - \langle x, w^* \rangle \end{aligned}$$

Note that by definition $g(w^*) = 0$, which is the optimal solution to the problem. In addition, an ε approximate solution to the ALR instance, assuming $\|w^*\| \leq 1$, corresponds to a vector $\hat{w}$ with $g(\hat{w}) \leq \varepsilon$. The following algorithm is an instantiation of Alg 3 from (Clarkson et al., 2012) applied to mathematical program 2 with the following arguments:

1. The primal decision set $\{\|w\| \leq 1\}$ and (linear) cost functions $c_x(w)$, admits an iterative low regret algorithm, namely online gradient descent, with expected regret $\mathbf{E}[R(T)] \leq 2\sqrt{T}$. This follows since the norms of x, w (for all $x \in \mathcal{K}$) are bounded by one. See e.g. Theorem 1 in (Zinkevich, 2003).

2. We assume oracle access to a procedure **Sample**$(\mathcal{K})$ that returns, for all $x \in \mathcal{K}$, an unbiased estimate of $c_x(w^*)$ with variance at most s.

The following theorem provides the analysis of Algorithm 2. Given the above it is immediately derived from Lemma 4.1 in (Clarkson et al., 2012).

Theorem 8. *Algorithm 2 runs in time $\tilde{O}(\frac{dn}{\varepsilon^2})$ and requires $O(\frac{d \log d \log(n)}{\varepsilon^2})$ queries to the procedure* **Sample**$(\mathcal{K})$. *It returns, with probability of at least $\frac{1}{2}$, a vector w such that $\max_{x \in \mathcal{K}} \langle w - w^*, x \rangle \leq \varepsilon$.*

In the following section we describe a validation procedure that can verify, w.h.p, whether a proposed hypothesis is correct. The validation procedure will not increase the asymptotic behavior of the sample nor running time complexity of the algorithm. With it, we can repeat the procedure of Theorem 8 $\log \frac{1}{\delta}$ many times, and achieve a high probability result described below.

Corollary 9. *There exists an algorithm that runs in time* $\tilde{O}(\frac{dn\log\frac{1}{\delta}}{\varepsilon^2})$ *and returns, with probability of at least* $1-\delta$*, a vector* w *such that* $\max_{x\in\mathcal{K}}\langle w-w^*,x\rangle\leq\varepsilon$.

Algorithm 2 Primal-Dual Algorithm

Input: T
Let $w_1\leftarrow 0$, $q_0\leftarrow \mathbf{1}_n$, $\eta\leftarrow\frac{1}{100}\sqrt{\frac{\log(n)}{T}}$.
for $t=1$ to T **do**
  Query **Sample**$(\mathcal{K})$ to obtains estimators $\tilde{a}_t(i)$ for all c_i's
  for $i\in[n]$ **do**
    $a_t(i)\leftarrow \text{clip}(\tilde{a}_t(i),1/\eta)$
    $q_t(i)\leftarrow q_{t-1}(i)(1-\eta a_t(i)+\eta^2 a_t(i)^2)$
  end for
  Choose $i_t\in[n]$ at random with $\Pr[i_t=i]\propto q_t(i)$
  $w_t\leftarrow w_{t-1}-\frac{1}{\sqrt{t}}\nabla_w c_{i_t}$, where $\nabla_w c_{i_t}=x_{i_t}$
end for
return $\bar{w}=\frac{1}{T}\sum_t w_t$

4.3. Validation

In this section we present Algorithm 3 that given a hypothesis w, verifies, w.h.p., that w is sufficiently accurate.

Algorithm 3 Verification

Input: Volumetric spanner S, parameters $\varepsilon,\delta>0$
run **Sample**$(\mathcal{K})$ $T=2\log(2n/\delta)|S|/\varepsilon^2$ times and obtain for each data point in $\mathcal{K}$, T i.i.d samples of $\langle w^*,x\rangle$
for each $x\in\mathcal{K}$ let $\tilde{f}(x)$ be the average of the above T samples.
declare w as accurate iff for all x, $|\langle w,x\rangle-\tilde{f}(x)|<2\varepsilon$

Lemma 10. *Algorithm 3 has the following properties: (1) it requires* $O(\log(n/\delta)|S|/\varepsilon^2)$ *queries to the oracle (2) if the worst-case error of* w *is bounded by* ε *then w.p. at least* $1-\delta$ *it will be declared as accurate (3) if the worst-case error of* w *is larger than* 3ε *then w.p. at least* $1-\delta$ *it will be declared as inaccurate.*

Proof. We recall that the process **Sample**$(\mathcal{K})$ returns unbiased estimates of $\langle w^*,x\rangle$ for all of the data points where the estimates are bounded in absolute value by $|S|$. By the Chernoff bound it can be verified that for any given $x\in\mathcal{K}$ it holds with probability at least $1-\delta/n$ that $|\langle w^*,x\rangle-\tilde{f}(x)|<\varepsilon$. A union bound shows that w.p. at least $1-\delta$ the above holds for all data points simultaneously. □

5. The Agnostic Case

In this section we discuss the agnostic case, where the learned function is not necessarily linear. The formal setting here is the following. There exist some function $f:\mathcal{K}\to[-1,1]$ for which we have noisy access. That is, a query at point $x\in\mathcal{K}$ returns an answer $\widehat{f(x)}$ consisting of $f(x)$ plus some zero mean noise. We denote by $w^*\in\mathbb{R}^d$ the optimal linear regressor, being the minimizer of the expression

$$\Phi(w^*)\triangleq\max_{x\in\mathcal{K}}|f(x)-\langle w^*,x\rangle|$$

We denote the attained value by λ. Our objective is to find a regressor w such that $\Phi(w)$ is as close to λ as possible. We analyze both an algorithmic result obtaining a multiplicative approximation to λ and proceed to show the tightness of this result in the sense that any algorithm achieving a better approximation must query all of the data points.

The following theorems describe both results correspondingly

Theorem 11. *Let* $f:\mathcal{K}\to[-1,1]$ *and let*

$$\lambda=\min_{w\in\mathbb{R}^d}\max_{x\in\mathcal{K}}|f(x)-\langle w,x\rangle|.$$

Given access to a noisy oracle of f*, algorithm 2 equipped with the* sample *procedure described in algorithm 1 outputs a regressor* w *such that with probability at least* $1-\delta$*, for each point* $x\in\mathcal{K}$ *it holds that*

$$|f(x)-\langle w,x\rangle|\leq O\left(\sqrt{d\log(d)}\cdot\lambda\right)+\varepsilon$$

The query complexity of the process is $O(d\log(n)\log(1/\delta)/\varepsilon^2)$.

Theorem 12. *For any integer* n, $0<\lambda<\ln(2n)/d$ *and any policy requiring* $o\left(n/(d\lambda^2)\right)$ *queries when given a set of* n *points in* $\mathbb{R}^d$*, there exist a set of points* $\mathcal{K}$ *of cardinality* $|\mathcal{K}|=n$ *and a function* $f:\mathcal{K}\to[-1,1]$ *with the following properties. First, there exists some* $w^*\in\mathbb{R}^d$ *such that* $\Phi(w^*)=\lambda$*. Second, the solution* w *obtained by the policy is such that*

$$\mathbf{E}[\Phi(w)]\geq\lambda\cdot c\sqrt{\frac{d}{\ln(2n)}}$$

where $c>0$ *is some universal constant. Here, the expectation is taken both over the (possible) internal randomness of the policy and the query noise.*

5.1. An Agnostic Algorithm

proof of Theorem 11. For simplicity we describe an algorithm with a success probability of at least $1/2$. The high probability result can be achieved analogously to the non-agnostic case (section 4.3). Recall from Algorithm 1 that S is the chosen volumetric spanner of $\mathcal{K}$ with $|S|=O(d)$. Recall that V is the $d\times|S|$ matrix whose columns are the elements of S, and $V^\dagger\in\mathbb{R}^{|S|\times d}$ is its Moore-Penrose pseudo

inverse. We also recall that that for all $x \in \mathcal{K}$ it holds for $\alpha_x = V^\dagger x$ that $\|\alpha_x\| \leq 1$ and $x = V\alpha_x$.

We define for $x \in \mathcal{K}$ the function $\tilde{f}(x) \triangleq f(V)^\top \alpha_x$ where $f(V)$ is the vector of length $|S|$ whose values are the values of f on the elements of S.

Observation 13. *$\tilde{f}(x)$ is a linear function. Furthermore, the procedure in algorithm 1 returns unbiased estimates of variance at most $|S|$ to $\tilde{f}$ at all points $x \in \mathcal{K}$ simultaneously.*

As a result of the above observation we get that by Theorem 8, when running Algorithm 2 with $T = \Theta(\log(n)|S|/\varepsilon^2)$ queries to f we obtain, with probability at least 1/2, a regressor w such that for all $x \in \mathcal{K}$,

$$\left|\langle w, x\rangle - \tilde{f}(x)\right| < \varepsilon \tag{2}$$

This is since all of our queries are made by uniformly sampling an element of S, hence the oracle we are querying is in fact an unbiased estimator for $\tilde{f}$. It thus remains to prove an upper bound for the ℓ_∞ distance between f and $\tilde{f}$. Let $w^* \in \mathbb{R}^d$ be the optimal regressor and let $b \in \mathbb{R}^n$ be the vector of biases corresponding to w^*; that is for $x \in \mathcal{K}$, $b(x) = \langle x, w^*\rangle - f(x)$. For $x \in \mathcal{K}$ we have that

$$\tilde{f}(x) = f(V)^\top \alpha_x = w^*(V)^\top \alpha_x + b(V)^\top \alpha_x =$$

$$\langle w^*, x\rangle + b(V)^\top \alpha_x$$

It follows that

$$\left|f(x) - \tilde{f}(x)\right| = \left|f(x) - \langle w^*, x\rangle - b(V)^\top \alpha_x\right| \overset{(1)}{\leq}$$

$$\lambda + \left|b(V)^\top \alpha_x\right| \overset{(2)}{\leq} \lambda + \|b(V)\| \overset{(3)}{\leq} \lambda\left(1 + \sqrt{|S|}\right) \tag{3}$$

Transitions (1) and (3) are due to the optimality of w^*. Inequality (2) is due to the Cauchy-Schwartz inequality and $\|\alpha_x\| \leq 1$. The proof follows from combining inequalities (2) and (3), along with $|S| = O(d)$.

□

5.2. Query Complexity Lower Bound

In order to prove the theorem we will present different scenarios that cannot be distinguished without using the required number of queries. To this end we will use a negative result for the PAC setting of the Multi-Armed Bandit (MAB) problem. We are interested in a special case of the problem; here, there exist some unknown vector $p \in [-1/2, 1/2]^n$ and the player can query indices of p. Each query is independent of the past and is a sample of a random variable in $[-1, 1]$ with expectation p_j where j is the index being queried. The goal of the player is to find an index j whose value p_j is maximal, i.e, find $\mathrm{maxarg}_j p_j$. The theorem below provides a lower bound to the number of queries required for the task. It is a special case of Theorem 5 in (Mannor & Tsitsiklis, 2004).

Lemma 14. *There exists a global constant c such that: For any $p \in [-1/2, 1/2]^n$ a policy that successfully identifies $\mathrm{maxarg}_j p_j$ with probability at least $2/3$, requires at least*

$$c\sum_{j \neq j^*} \frac{1}{(p_{j^*} - p_j)^2}$$

queries, where $j^* = \mathrm{maxarg}_j p_j$.

We begin with the construction of the set of data points. Let $x_1, \ldots, x_n$ be i.i.d random points chosen uniformly from the normalized hypercube $\{-1/\sqrt{d}, 1/\sqrt{d}\}^d$. We require the points to be uncorrelated in a manner guaranteed by the following lemma

Lemma 15. *With probability larger than $1/2$, it holds that for every $j \neq j' \in [n]$*

$$|\langle x_j, x_{j'}\rangle| \leq \sqrt{4\ln(2n)/d}$$

Proof. Notice that $\langle x_j, x_{j'}\rangle = \sum_{i=1}^d (x_j)_i \cdot (x_{j'})_i = \frac{1}{d}\sum_{i=1}^d Y_i$ where the Y_i's are i.i.d r.v chosen uniformly from $\{-1, 1\}$. By the well known Chernoff inequality we get that for any $t > 0$

$$\Pr\left[\left|\sum_{i=1}^d Y_i\right| > t\sqrt{d}\right] \leq 2\exp(-t^2/2)$$

Taking $t = \sqrt{4\ln(2n)}$ ensures the above probability is at most $1/n^2$ and a union bound argument proves the lemma. □

We define the set of data points as follows: $x_1, \ldots, x_n$ are points in the normalized hypercube for which the statement of the above lemma applies. We define n scenarios, enumerated from 1 to n that will be indistinguishable by the policy. For $j \in [n]$, scenario j is defined as follows: $f(x_{j'}) = 0$ for all $j' \in [n]$ where $j' \neq j$ and $f(x_j) = \lambda\sqrt{d/4\ln(2n)}$.

We first analyze the regressors of f in the different scenarios.

Lemma 16. *In any of the above scenarios, there exists a linear regressor w^* such that $|f(x_j) - \langle w^*, x_j\rangle| \leq \lambda$ for all j. Furthermore, any regressor will achieve a worst case error of at least $\sqrt{d/16\ln(2n)}\lambda$ in $n-1$ of the n scenarios.*

Proof. We prove the first claim for scenario 1 and notice that the other scenarios have analogous proofs. Let $w =$

$x_1 \cdot \lambda \cdot \sqrt{d/4\ln(2n)}$. Notice that due to our assumption on λ, w is contained in the unit sphere. For x_1, we have $f(x_1) = \langle w, x_1 \rangle$. For $j > 1$ we have

$$|f(x_j) - \langle w, x_j \rangle| = |\langle w, x_j \rangle| =$$

$$|\langle x_1, x_j \rangle| \cdot \lambda \cdot \sqrt{d/4\ln(2n)} \leq \lambda$$

This proves the first claim in the lemma.

To prove the second claim, we define w as an arbitrary regressor and consider two cases. In one, $\langle w, x_j \rangle < \sqrt{d/16\ln(2n)}\lambda$ for all j. Clearly, the worst case error is $\sqrt{d/16\ln(2n)}\lambda$ in all scenarios. In the other case, for some point j, $\langle w, x_j \rangle \geq \sqrt{d/16\ln(2n)}\lambda$. In any scenario except scenario j, $f(x_j) = 0$ hence the worst case error is at least $\sqrt{d/16\ln(2n)}\lambda$, proving the claim. □

We now conclude the proof of Theorem 12 with a lemma showing that without many queries, it is not possible to identify the correct scenario.

Lemma 17. *Any policy that identifies the correct scenario with probability at least $2/3$ requires $\Omega(n/(d\lambda^2))$ queries.*

Proof. Consider the vector $p \in \mathbb{R}^n$ where $p_j = f(x_j)$. Assuming sufficiently small λ, $p \in [-1/2, 1/2]^n$. Also, for $j^* = \text{maxarg}_j p_j$, it holds that

$$\sum_{j \neq j^*} \frac{1}{(p_{j^*} - p_j)^2} = \Omega(n/(d\lambda^2))$$

Finnaly, a policy that identifies the scenario with probability at least $2/3$ implicitly identifies the index j^* with the same probability and by that solves the exploratory MAB problem associated with p. Hence, by Lemma 14, it must use

$$\Omega\left(\sum_{j \neq j^*} \frac{1}{(p_{j^*} - p_j)^2}\right) = \Omega\left(n/(d\lambda^2)\right)$$

queries. □

proof of Theorem 12. Assume that the scenario defining f is chosen uniformly at random among the n scenarios presented above. Consider a policy using $o(n/(d\lambda^2))$ queries. According to Lemma 17, the correct scenario cannot be identified with probability larger than $2/3$. This, and Lemma 16 indicate that with probability at least $1/3$ the regressor chosen will have a worst-case additive error of at least $\sqrt{d/16\ln(2n)}\lambda$. The claim of the theorem now follows. □

6. Conclusions and Open Questions

We have studied active real-valued learning (i.e. regression) in the presence of noise, which has been studied before in the experiment-design literature. While exponential separation of active and passive learning is notoriously hard, in the fundamental hard-margin linear regression setting we show it to be attainable: while passive algorithms require $O(\frac{n}{\varepsilon^2})$ queries in order to correctly deduce the optimal linear regressor up to precision ε, our active algorithm requires only $\tilde{O}(\frac{d}{\varepsilon^2})$ queries.

The latter is attainable using recently-developed techniques in sublinear optimization and exploration using volumetric ellipsoids.

We continue to study active regression in the agnostic setting, where we show that the same algorithm gives a $\sqrt{d}$-approximation to the optimal linear regressor, and that no algorithm can do better in general without querying all data points.

Many intriguing questions remain open: what is the query complexity of active non-linear regression? Of particular interest is the characterization of the query complexity of the most popular regression models, i.e. polynomial and support-vector regression.

Acknowledgements

The first author gratefully acknowledges support by the Microsoft-Technion EC center, ISF grant 810/11 and an ERC Starting Grant project SUBLRN.

References

Atkinson, A.A.C. and Donev, A.A.N. *Optimum Experimental Designs*. Oxford science publications. OXFORD University Press, 1992. ISBN 9780198522546. URL http://books.google.co.il/books?id=cmmOA_-M7S0C.

Audibert, Jean-Yves and Catoni, Olivier. Linear regression through PAC-Bayesian truncation. 2010. URL http://hal.archives-ouvertes.fr/hal-00522536.

Audibert, Jean-Yves and Catoni, Olivier. Robust linear least squares regression. *Annals of Statistics*, 2011.

Balcan, Maria-Florina, Beygelzimer, Alina, and Langford, John. Agnostic active learning. *J. Comput. Syst. Sci.*, 75(1):78–89, January 2009. ISSN 0022-0000. doi: 10.1016/j.jcss.2008.07.003. URL http://dx.doi.org/10.1016/j.jcss.2008.07.003.

Clarkson, Kenneth L., Hazan, Elad, and Woodruff, David P.

Sublinear optimization for machine learning. *J. ACM*, 59 (5):23:1–23:49, November 2012.

Cohn, David, Atlas, Les, and Ladner, Richard. Improving generalization with active learning. *Mach. Learn.*, 15(2): 201–221, May 1994. ISSN 0885-6125. doi: 10.1023/A:1022673506211. URL http://dx.doi.org/10.1023/A:1022673506211.

Dasgupta, Sanjoy and Langford, John. Active learning tutorial. *Proceedings of the 26th International Conference on Machine Learning (ICML)*, 2009. URL http://hunch.net/~active_learning/.

Dasgupta, Sanjoy, Kalai, Adam Tauman, and Monteleoni, Claire. Analysis of perceptron-based active learning. *Journal of Machine Learning Research*, 10:281–299, 2009.

Ganti, Ravi and Gray, Alexander G. Upal: Unbiased pool based active learning. In *AISTATS*, volume 22 of *JMLR Proceedings*, pp. 422–431. JMLR.org, 2012.

Györfi, L. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer, 2002. ISBN 9780387954417. URL http://books.google.co.il/books?id=Q9NMAp-cKy0C.

Hanneke, Steve. Teaching dimension and the complexity of active learning. In *Proceedings of the 20th annual conference on Learning theory*, COLT'07, pp. 66–81, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 978-3-540-72925-9. URL http://dl.acm.org/citation.cfm?id=1768841.1768851.

Hazan, Elad, Karnin, Zohar, and Mehka, Raghu. Volumetric spanners and their applications to machine learning. *CoRR*, abs/1312.6214, 2013.

Hsu, Daniel, Kakade, Sham M., and Zhang, Tong. Random design analysis of ridge regression. *Journal of Machine Learning Research - Proceedings Track*, 23:9.1–9.24, 2012.

Mannor, Shie and Tsitsiklis, John N. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

McCallum, Andrew and Nigam, Kamal. Employing em and pool-based active learning for text classification. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML)*, pp. 350–358, 1998.

Spruill, MC and Studden, WJ. A kiefer-wolfowitz theorem in a stochastic process setting. *The Annals of Statistics*, 7(6):1329–1332, 1979.

Vapnik, Vladimir. *The nature of statistical learning theory*. springer, 2000.

Wu, Chien-Fu. Some algorithmic aspects of the theory of optimal designs. *Annals of Statistics*, 6(6):1286–1301, 1978.

Zinkevich, Martin. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, pp. 928–936, 2003.