Journal of Machine Learning Research 20 (2019) 1-23 Submitted 9/17; Revised 12/18; Published 1/19

An Approach to One-Bit Compressed Sensing Based on Probably Approximately Correct Learning Theory

Mehmet Eren Ahsen MEHMETEREN.AHSEN@MSSM.EDU
Icahn School of Medicine at Mount Sinai
1468 Madison Avenue
New York, NY 10029

Mathukumalli Vidyasagar M.VIDYASAGAR@UTDALLAS.EDU,M.VIDYASAGAR@IITH.AC.IN
Department of Systems Engineering
University of Texas at Dallas
Richardson, TX 75080, USA, and
Department of Electrical Engineering
Indian Institute of Technology Hyderabad
Kandi, Telangana, India 502285

Editor: John Shawe-Taylor

Abstract

In this paper, the problem of one-bit compressed sensing (OBCS) is formulated as a problem in probably approximately correct (PAC) learning. It is shown that the Vapnik-Chervonenkis (VC-) dimension of the set of half-spaces in $\mathbb{R}^n$ generated by k-sparse vectors is bounded below by $k(\lfloor \lg(n/k) \rfloor + 1)$ and above by $\lfloor 2k \lg(en) \rfloor$. By coupling this estimate with well-established results in PAC learning theory, we show that a consistent algorithm can recover a k-sparse vector with $O(k \lg n)$ measurements, given only the signs of the measurement vector. This result holds for *all* probability measures on $\mathbb{R}^n$. The theory is also applicable to the case of noisy labels, where the signs of the measurements are flipped with some unknown probability.

1. Introduction

The field of "compressed sensing" has become very popular in recent years, with an explosion in the number of papers. Stated briefly, the core problem in compressed sensing is to recover a high-dimensional sparse (or nearly sparse) vector x from a small number of measurements of x. In the traditional problem formulation, the measurements are *linear*, consisting of m real numbers $y_i = \langle a_i, x \rangle, i = 1, \ldots, m$, where the measurement vectors $a_i \in \mathbb{R}^n$ are chosen by the learner. More recently, attention has focused on so-called one-bit compressed sensing, referred to hereafter as OBCS in the interest of brevity. In OBCS the measurements consist, not of the inner products $\langle a_i, x \rangle$, but rather *just the signs of these inner products*, i.e., a "one-bit" representation of these measurements. In much of the OBCS literature, the vectors a_i are chosen at random from some specified probability distribution, often Gaussian. Because the sign of $\langle a_i, x \rangle$ is unchanged if x is replaced by any positive multiple of x, it is obvious that under this model, one can at best aspire to recover the unknown

vector x only to within a positive multiple, or equivalently, to recover the normalized vector $x/\|x\|_2$. This limitation can be overcome by choosing the measurements to consist of the signs of inner products $\langle a_i, x\rangle + b_i$, where again a_i, b_i are selected at random.

The current status of OBCS is that while several algorithms have been proposed, theoretical analysis is available for only a few algorithms. Moreover, in cases where theoretical analysis is available, the sample complexity is extremely high. In the present paper, we interpret OBCS as a problem in probably approximately correct (PAC) learning theory, which is a well-established branch of statistical learning theory. By doing so, we are able to draw from the wealth of results that are already available, and thereby address some of the currently outstanding issues. In PAC learning theory, a central role is played by the so-called Vapnik-Chervonenkis (VC) dimension of the collection of concepts to be learned. The principal result of the present paper is that the VC-dimension of the set of half-planes in $\mathbb{R}^n$ generated by k-sparse vectors is bounded below by $k\lg(n/k)$ and above by $2k\lg(en)$, plus some roundoff terms. Using this bound, we are able to establish the following results for the case where $x \in \mathbb{R}^n$ has no more than k nonzero components:

- In principle, OBCS is possible whenever the measurement vector (a_i, b_i) is drawn at random from *any arbitrary* probability distribution. Moreover, if a consistent algorithm[1] can be devised, then the number of measurements is $O(k \lg n)$.
- There is also a *lower bound* on the OBCS problem. Specifically, there exists a probability distribution on $\mathbb{R}^n$ such that, if (a_i, b_i) is drawn at random from this distribution, then the number of measurements required to learn each k-sparse n-dimensional vector is bounded *below* by $\Omega(k\ln(n/k))$.
- It is shown that OBCS is possible under random flipping of the signs of the measurements. This finding builds on earlier results on PAC learning with noisy labels.
- If the samples a_i are not independent, but are β-mixing, learning is still possible, and explicit estimates are available for the rate of learning.

The paper is organized as follows: In Section 2, a brief review is given of some recent papers in OBCS. In Section 3, some parts of PAC learning theory that are relevant to OBCS are reviewed. In particular, it is shown how OBCS can be formulated as a problem in PAC learning, so that OBCS can be addressed by finding upper bounds on the VC-dimension of half-spaces generated by k-sparse vectors. In Section 4, both upper and lower bounds are derived for the VC-dimension of half-spaces generated by k-sparse vectors. In Section 5, the case where the sign measurements are flipped at random is examined. Finally, Section 6 contains a discussion of some issues that merit further investigation.

2. Brief Review of One-Bit Compressed Sensing

By now there is a substantial literature regarding the traditional compressed sensing formulation, out of which only a few references are cited here in the interests of brevity. Book-length treatments of compressed sensing can be found in (Elad, 2010; Foucart and Rauhut, 2013; Rish and Grabarnik, 2015; Hastie et al., 2015). Amongst these, (Foucart

1. This term is standard in statistical learning theory and is defined later.

and Rauhut, 2013) contains a thorough discussion of virtually all aspects of compressed sensing theory. A recent volume (Eldar and Kutyniok, 2012) is a compendium of articles on a variety of topics. The first paper in this volume (Davenport et al., 2012) is a survey of the basic results in compressed sensing. Another paper (Negabhan et al., 2012) provides a very general framework for sparse regression that can be used, among other things, to analyze compressed sensing algorithms. Each of these papers contains an extensive bibliography.

Throughout this paper, n denotes some fixed and large integer. For $x \in \mathbb{R}^n$, let $\text{supp}(x)$ denote the support of a vector, and let Σ_k denote the set of k-sparse vectors in $\mathbb{R}^n$; that is,

$$\text{supp}(x) := \{i : x_i \neq 0\}, \Sigma_k := \{x \in \mathbb{R}^n : |\text{supp}(x)| \leq k\}.$$

Suppose $x \in \mathbb{R}^n$ is k-sparse, that is, $x \in \Sigma_k$, where both n and k are known integers with $k \ll n$. The basic problem in compressed sensing is to design an $m \times n$ matrix A where $m \ll n$, together with a decoder map $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ such that $\Delta(Ax) = x$ for all $x \in \Sigma_k$, that is, x can be recovered exactly from the m-dimensional vector of linear measurements $y = Ax$. Variations of the problem include the case where x is not exactly sparse, and/or $y = Ax + \eta$ where η is a measurement noise. Given a vector $x \in \mathbb{R}^n$, an integer $k < n$, define

$$\sigma_k(x, \|\cdot\|_1) := \min_{z \in \Sigma_k} \|x - z\|_1$$

to be the **k-sparsity index** of the vector x. It is clear that $x \in \Sigma_k$ if and only if $\sigma_k(x, \|\cdot\|_1) = 0$. By far the most popular method for recovering a sparse vector is ℓ_1-norm minimization. If $y = Ax + \eta$, the approach is to define the "decoder" map $\Delta : \mathbb{R}^m \to \mathbb{R}^n$ by

$$\Delta(y) = \hat{x} := \underset{z}{\text{argmin}} \|z\|_1 \text{ s.t. } \|y - Az\|_2 \leq \epsilon, \tag{1}$$

where ϵ is a known upper bound on $\|\eta\|_2$. In a series of papers by Candès, Tao, Donoho, and others, it is demonstrated that if the matrix A is chosen so as to satisfy the so-called Restricted Isometry Property (RIP), then the decoder Δ defined in (1) produces a good approximation to x in the following sense: There exist constants C, D that depend on the measurement matrix A but not on the unknown vector x, such that

$$\|\hat{x} - x\|_1 \leq C\sigma_k(x, \|\cdot\|_1) + D\epsilon.$$

In particular, ℓ_1-norm minimization recovers x exactly if $x \in \Sigma_k$ and $\eta = 0$. See for example (Candès and Tao, 2005; Candès et al., 2006; Donoho, 2006a,b), as well as the survey paper (Davenport et al., 2012) and the comprehensive book (Foucart and Rauhut, 2013). Further, it is shown in (Candès and Tao, 2005) that if the elements of A are samples of independent and identically distributed (i.i.d.) normal random variables (denoted by $a_{ij} \sim \mathcal{N}(0, 1)$), then the resulting normalized matrix $(1/\sqrt{m})A$ satisfies the RIP with probability that can be made arbitrarily close to one by increasing the number of measurements.

The remainder of the section is devoted to a discussion of the one-bit compressed sensing (OBCS) problem. One bit compressed sensing is introduced in (Boufounos and Baraniuk, 2008). In that paper, it is assumed that the measurement y_i equals the bipolar quantity $y_i = \text{sign}(\langle a_i, x \rangle)$, as opposed to the real number $\langle a_i, x \rangle$. Because the measurements remain invariant if x is replaced by any positive multiple of x, there is no loss of generality in

assuming that $\|x\|_2 = 1$. A greedy algorithm called "renormalized fixed point iteration" is introduced, as follows:

$$\hat{x} := \operatorname*{argmin}_z \|z\|_1 + \lambda \sum_{i=1}^{m} f(|(YA)_i|) \text{ s.t. } \|z\|_2 = 1,$$

where $y_1, \ldots, y_m$ are some constants, $Y = \text{Diag}(y_1, \ldots, y_m)$, and the regularizing function $f(\cdot)$ is defined by

$$f(\alpha) := \begin{cases} \alpha^2/2 & \text{if } \alpha < 0, \\ 0 & \text{if } \alpha \geq 0. \end{cases}$$

The optimization problem is non-convex due to the constraint $\|z\|_2 = 1$. Only simulations are provided, but no theoretical results.

In (Gupta et al., 2010), the focus is on recovering the support set of the unknown vector x from noise-corrupted measurements of the form $\text{sign}(\langle a_i, x\rangle + \eta_i)$, where the noise vector $\boldsymbol{\eta}$ consists of pairwise independent Gaussian signals. A non-adaptive algorithm is presented that makes use of Hoeffding's inequality applied to the expected value of the covariance of the signs of two Gaussian random variables. An adaptive algorithm is also presented. In (Boufounos, 2009), a new greedy algorithm is presented called "matched signed pursuit." The optimization problem is not convex; as a result there are no theoretical resuts. The algorithm is similar to the CoSaMP algorithm for the conventional compressed sensing problem (Needell and Tropp, 2009).

In (Jacques et al., 2013), the authors introduce a constant

$$\epsilon_{\text{opt}} := \frac{k}{2em + 2k^{3/2}},$$

where e denotes the base of the natural logarithm, and show that it is the minimum achievable error no matter what reconstruction method is used. Next, they choose any $\epsilon \geq \epsilon_{\text{opt}}$ and show the following result: Let $A \sim \mathcal{N}^{m \times n}(0, 1)$ consist of mn pairwise independent normal random variables, and let $y_i = \text{sign}(\langle a_i, x\rangle)$. Fix $\delta \in (0, 1)$. If the number of measurements m satisfies

$$m \geq \frac{2}{\epsilon}\left(2k \ln n + 4k \ln \frac{17}{\epsilon} + \ln \frac{1}{\delta}\right).$$

then for every pair $x, s \in \Sigma_k$,

$$\text{sign}(Ax) = \text{sign}(As) \implies \|x - s\|_2 \leq \epsilon,$$

with probability $\geq 1 - \delta$. In words, this result means that if we can find a k-sparse vector s that is consistent with the observation vector y, then s is close to x. In fact, s can be made as close to x as desired by increasing the number of measurements m. Unfortunately, this result is not practical because finding such a vector s is equivalent to finding a minimal ℓ_0-norm solution consistent with the observations, which is known to be an NP-hard problem (Natarajan, 1995).

In (Plan and Vershynin, 2013a), the authors focus on vectors $x \in \mathbb{R}^n$ that satisfy an inequality of the form $\|x\|_1/\|x\|_2 \leq s$. Note that if x is s-sparse, then it satisfies the above

inequality, though of course the converse is not true. Thus they use the ratio $\|x\|_1/\|x\|_2$ as a proxy for $\|x\|_0$. They choose measurement vectors $a_i \in \mathbb{R}^n$ according to the Gaussian distribution, or more generally, any radially invariant distribution; this means that, under the chosen probability distribution on the vector $a \in \mathbb{R}^n$, the normalized vector $a/\|a\|_2$ is uniformly distributed on the sphere $S^{n-1} \subseteq \mathbb{R}^n$. With these randomly generated measurement vectors, the measured quantities are $y_i = \text{sign}(\langle a_i, x\rangle)$. The authors propose to estimate x via

$$\hat{x} := \underset{z}{\text{argmin}} \|z\|_1 \text{ s.t. } \text{sign}(\langle a_i, z\rangle) = y_i \; \forall i, \sum_{i=1}^{m} |\langle a_i, z\rangle| = m, \tag{2}$$

where m is the number of measurements. They show that if

$$\epsilon > C\left(\frac{s}{m}\ln(2n/s)\ln(2n/m + 2m/n)\right)^{1/5}$$

for some universal constant C, then with probability $\geq 1 - \exp(-c\epsilon m)$ where c is another universal constant, it is true that

$$\left\| \frac{x}{\|x\|_2} - \frac{\hat{x}}{\|\hat{x}\|_2} \right\|_2 \leq \epsilon \tag{3}$$

for all $x \in \mathbb{R}^n$ such that $\|x\|_1/\|x\|_2 \leq \sqrt{s}$. Although this is the first proposed convex algorithm to recover x, the number of measurement m is $O(\delta^{-5})$. From (3) we see that if we are able to carry out the ℓ_0-norm minimization, then m is $O(\delta^{-1})$. It is still an open question whether or not a practical algorithm can achieve this optimal dependence on δ in (3). In (Ai et al., 2014) the theory is extended to non-Gaussian noise signals that are sub-Gaussian.

In (Plan and Vershynin, 2013b), it is assumed that the measurements $y_i \in \{0, 1\}$ are drawn at random with the property that the expected value $E(y_i)$ satisfies

$$E(y_i) = \theta(\langle a_i, x\rangle),$$

where $\theta : \mathbb{R} \to [-1, 1]$ is an unknown function. If $\theta(\alpha) = \tanh(\alpha/2)$, then the problem is one of logistic regression, whereas if $\theta(\alpha) = \text{sign}(\alpha)$, then the problem becomes OBCS. A probabilistic approach is proposed in (Plan and Vershynin, 2013b), which has the advantage that the resulting optimization problem is convex. However, the disadvantage is that the number of measurements m is $O(\delta^{-6})$ where δ is the probability that the algorithm may fail. The large negative exponent of δ makes the algorithm somewhat impractical.

In all of the papers discussed until now, the measurement vector y_i equals $\text{sign}(\langle a_i, x\rangle)$ for suitably generated random vectors a_i. As mentioned above, with such a set of measurements one can at best aspire to recover only the normalized unknown vector $x/\|x\|_2$. In (Knudson et al., 2016), it is proposed to overcome this limitation by changing the *linear* measurements to *affine* measurements. Specifically, the measurements in (Knudson et al., 2016) are of the form $y_i = \text{sign}(\langle a_i, x\rangle + b_i)$, where $a_{ij} \sim \mathcal{N}(0, 1)$,[2] and $b_i \sim \mathcal{N}(0, \tau^2)$ where τ is some

2. Recall that each a_i is an n-vector.

specified constant. If a prior upper bound R for $\|x\|_2$ is available, then it is possible to choose $\tau = R$. Then the optimization problem in (2) is modified to[3]

$$(\hat{x}, \hat{v}) := \underset{z,v}{\operatorname{argmin}}[\|z\|_1 + \tau|v|] \text{ s.t. } \operatorname{sign}(\langle a_i, z\rangle + b_i v) = y_i \ \forall i, \sum_{i=1}^{m} |\langle a_i, z\rangle + b_i v| = m. \quad (4)$$

It is evident that the above formulation is similar to the formulation in (Plan and Vershynin, 2013a) applied to the augmented vector $(x, v) \in \mathbb{R}^{n+1}$. The following result is shown in (Knudson et al., 2016, Theorem 4): Fix τ, R, α such that $\alpha < \min\{1, \tau/2\}$. If

$$m \geq C \left(\frac{\sqrt{R^2 + \tau^2}}{\alpha} \right)^5 \log^2 \left(\frac{2n}{x} \right),$$

then for all vectors $x \in \mathbb{R}^n$ with $\|x\|_1/\|s\|_2 \leq \sqrt{s}$, the solution $(\hat{x}, \hat{v})$ to the optimization problem in (4) satisfies

$$\|(\hat{x}/\hat{v}) - x\|_2 \leq \frac{4\sqrt{R^2 + \tau^2}}{\tau}\alpha,$$

with a probability exceeding

$$1 - C \exp\left(-\frac{c\alpha m}{\sqrt{R^2 + \tau^2}} \right),$$

where C and c are universal constants. If R is a known prior upper bound for $\|x\|_2$, then one can choose $\tau = R$ in the above, in which the bound simplifies to

$$\|(\hat{x}/\hat{v}) - x\|_2 \leq 4\sqrt{2}\alpha.$$

3. Preliminaries

In this section we present some preliminary results, while the main results are presented in the next section. As shown below, the one-bit compressed sensing (OBCS) problem can be naturally formulated as a problem in probably approximately correct (PAC) learning. In fact, several of the approaches proposed thus far for solving the OBCS problem are similar to existing methods in PAC learning, but do not take full advantage of the power and generality of PAC learning theory. Some of the things that "come for free" in PAC learning theory are: explicit estimates for the number of measurements m, ready extension to the case where successive measurement vectors a_i are not independent but form a β-mixing process, and ready extension to the case of noisy measurements. The negative is that the PAC learning approach does not always address the *efficiently computability* of the algorithms. In (Valiant, 1984) which launched PAC learning theory, efficient computability of the algorithm is a central consideration. However, subsequent literature does not always pay full attention to this aspect.

3. Note that v here equals u/τ in (Knudson et al., 2016, Eq. (6)). Also, since δ is used to denote the confidence of a PAC learning algorithm later in the paper, we use α instead of δ as in the cited equation.

3.1. Brief Introduction to the PAC Learning Problem

In this subsection, we give a brief introduction to PAC learning theory. By now the fundamentals of PAC learning theory are well-developed, and several book-length treatments are available, including (Vapnik, 1998; Anthony and Bartlett, 1999; Vidyasagar, 1997, 2003). The theory encompasses a wide variety of learning situations. However, OBCS is aligned closely with the most basic version of PAC learning, known as concept learning, which is formally described next.

The concept learnig problem is formulated over a metric space X (usually taken as $\mathbb{R}^n$ for some integer n), together with the associated Borel σ-algebra $\mathcal{S}$; a collection of sets $\mathcal{C} \subseteq \mathcal{S}$, known as the **concept class**; and a family $\mathcal{P}$ of probability measures on $(X, \mathcal{S})$. If $\mathcal{P} = \mathcal{P}^*$, the set of *all* probability measures on $(X, \mathcal{S})$, the problem is known as one of "distribution-free learning."

Learning takes place as follows: A fixed but unknown set $T \in \mathcal{C}$, known as the "target concept," is chosen, along with an unknown probability measure $P \in \mathcal{P}$. Then random samples $\{c_1, c_2, \ldots\}$ are generated independently in accordance with the chosen distribution P. This is the basic version of PAC learning studied in (Vapnik, 1998; Anthony and Bartlett, 1999; Vidyasagar, 1997). The case where the sample sequence can exhibit dependence, for example if they come from a Markov process with the stationary distribution P, is studied in (Vidyasagar, 2003). Using the sample c_i, an "oracle" generates a "label" $y_i \in \{0, 1\}$. In the case of noise-free measurements, $y_i = I_T(c_i)$, where T is the fixed but unknown target concept, and $I_T(\cdot)$ denotes the indicator function of the set T. In the case of noisy measurements, the label y_i equals $I_T(c_i)$ flipped with an unknown probability α. This situation is studied in Section 5. After m such samples are drawn and labelled, the available information $\{(c_i, y_i)\}_{i=1}^m \in (X \times \{0, 1\})^m$ is passed through an "algorithm" to generate a "hypothesis," or an approximation to the unknown target concept T. For present purposes, an "algorithm" is any indexed collection of maps $\{A_m\}_{m \geq 1}$ where

$$A_m : (X \times \{0, 1\})^m \to \mathcal{C}.$$

In other words, an algorithm is any systematic procedure for taking a finite sequence of labelled samples, and returning an element of the concept class $\mathcal{C}$. The concept

$$G_m(T; \mathbf{c}) := A_m \left(\{(c_i, y_i)\}_{i=1}^m \right)$$

is called the "hypothesis" generated by the first m samples when the sample sequence is $\mathbf{c} = (c_1, \ldots, c_m)$, and the label sequence is $\mathbf{y} = (y_1, \ldots, y_m)$. In the interests of reducing clutter, we will use G_m in the place of $G_m(T; \mathbf{c})$ unless the full form is needed for clarity. Note that A_m is a deterministic map, but the hypothesis G_m is random because it depends on the random learning sequence $\{c_i\}$.

To measure how well the hypothesis G_m approximates the unknown target concept T, we use the **generalization error** defined by

$$J(T, G_m) = E[|I_T(x) - I_{G_m}(x)|, P]. \tag{5}$$

Thus $J(T, G_m)$ is the expected value of the difference between the indicator function $I_T(\cdot)$ and the label generated by the oracle with the input $I_{G_m}(\cdot)$. Note that both I_T and I_{G_m}

assume values in $\{0,1\}$. Hence $J(T,G_m)$ is also the probability that, when a random test sample $x \in X$ is generated in accordance with the probability distribution P, the sample is *misclassified* by the hypothesis G_m, in the sense that $I_T(x) \neq I_{G_m}(x)$. In particular, if the hypothesis G_m differs from the target concept T by a set of measure zero, then the generalization error would be zero.

The key quantity in PAC learning theory is the **learning rate**, defined by

$$r(m,\epsilon) := \sup_{P\in\mathcal{P}} \sup_{T\in\mathcal{C}} P^m\{\mathbf{c} \in X^m : J(T,G_m) > \epsilon\}. \tag{6}$$

Therefore $r(m,\epsilon)$ is the worst-case measure, over all probability distributions in $\mathcal{P}$ and all target concepts in $\mathcal{C}$, of the set of "bad" samples $\mathbf{c} = (c_1,\ldots,c_m)$ for which the corresponding hypothesis G_m has a generalization error larger than a prespecifie threshold ϵ. Thus, after m samples are generated together with their labels, and the hypothesis G_m is generated using the algorithm, it can be asserted with confidence $1 - r(m,\epsilon)$ that G_m will correctly classify the next randomly generated test sample with probability of at least $1-\epsilon$.

Definition 1 *An algorithm $\{A_m\}$ is said to be* **probably approximately correct (PAC)** *if $r(m,\epsilon) \to 0$ as $m \to \infty$, for every fixed $\epsilon > 0$. The concept class $\mathcal{C}$ is said to be* **PAC learnable** *under the family of probability measures $\mathcal{P}$ if there exists a PAC algorithm.*

The objective of statistical learning theory is to determine conditions under which there exists a PAC algorithm for a given concept class, and if so, to find upper bounds for the learning rate $r(m,\epsilon)$.

3.2. OBCS as a Problem in PAC Learning

In order to embed the problem of one-bit compressed sensing into the framework of concept learning, we proceed as follows. We begin with the case where the measurements are of the form $\text{sign}(\langle a_i, x\rangle)$ where the a_i are chosen at random according to some arbitrary probability distribution, *which need not be the Gaussian.* Observe now that the closed half-space

$$H(x) := \{z \in \mathbb{R}^n : \langle z, x\rangle \geq 0\}$$

determines the vector x uniquely to within a positive scalar multiple, because x is the normal to the half-space. Conversely, the vector x uniquely determines the corresponding half-space $H(x)$, which remains invariant if x is replaced by a positive multiple of x. Thus the OBCS problem can be posed as that of determining the half-space $H(x)$ given the measurements $\text{sign}(\langle a_i, x\rangle), i = 1,\ldots,m$ where the a_i are selected at random in accordance with some probability measure P. Moreover, the one-bit measurement $\text{sign}(\langle a_i, x\rangle)$ equals $2I_{H(x)}(a_i) - 1$, where $I_{H(x)}(\cdot)$ denotes the indicator function of the half-space $H(x)$. Therefore, to within a simple affine transformation, the OBCS problem becomes that of determining an unknown half-space $H(x)$ from labelled samples $(a_i, I_{H(x)}(a_i)), i = 1,\ldots,m$, where the samples a_i are generated at random according to some prespecified probability measure. This is a PAC learning problem where the underlying space X is $\mathbb{R}^n$, with $\mathcal{S}$ being the associated Borel σ-algebra. The concept class $\mathcal{C}$ is the collection of all half-spaces $\{H_k^n(x)\}$ where

$$H_k^n(x) = \{a \in \mathbb{R}^n : \langle a, x\rangle \geq 0\} \tag{7}$$

as x varies over Σ_k, the set of k-sparse vectors in $\mathbb{R}^n$. The family of probability measures $\mathcal{P}$ can either be a singleton $\{P\}$ where P is specified *a priori*, or $\mathcal{P}^*$, the family of *all* probability measures on $\mathbb{R}^n$, or anything in-between.

When measurements are of the type $\text{sign}(\langle a_i, x\rangle)$, it is inherently impossible to determine the unknown vector x, except to within a positive scalar multiple. This is addressed by changing the measurements to be of the form $\text{sign}(\langle a_i, x\rangle + b_i)$, as suggested in (Knudson et al., 2016). Some slight modifications are required to address this modified formulation. In this case theunderlying space X is $\mathbb{R}^{n+1}$, and $\mathcal{S}$ is the associated σ-algebra. The concept class $\mathcal{C}$ is the collection of all half-spaces $\{H_k^{n+1}\}$ where

$$H_k^{n+1}(x) = \{(a, b) \in \mathbb{R}^{n+1} : \langle a, x\rangle + b \geq 0\} \tag{8}$$

as x varies over Σ_k. Finally, the family of probability measures $\mathcal{P}$ can either be a singleton $\{P\}$ where P is specified *a priori*, or $\mathcal{P}^*$, the family of *all* probability measures on $\mathbb{R}^n$, or anything in-between.

For a given $x \in \Sigma_k$, if $a \in \mathbb{R}^n$ belongs to the half-space $H_k^n(x)$, then the vector $(a, 0) \in \mathbb{R}^{n+1}$ belongs to $H_k^{n+1}(x)$. However, the half-space $H_k^{n+1}(x)$ can also contain vectors of the form (a, b) with $b \neq 0$.

3.3. Interpretation of the Generalization Error

Suppose P is a probability measure on $(X, \mathcal{S})$, and define the quantity

$$d_P(A, B) := P(A\Delta B),\ \forall A, B \in \mathcal{S}.$$

Then d_P is a pseudometric on $\mathcal{S}$, in that d_P satisfies all the axioms of a metric, except that $d_P(A\Delta B) = 0$ does not necessarily imply that $A = B$, only that $A\Delta B$ has measure zero under P. In the traditional PAC learning problem formulation, the quantity of interest is the generalization error defined in (5). It is easy to see that an alternate expression for the generalization error is

$$J(T, G_m) = P(A\Delta B) = d_P(A, B).$$

Therefore the traditional PAC learning formulation does not distinguish between two concept classes that differ by a set of measure zero.

The above discussion explains the limitations of one-bit compressed sensing as described in (Plan and Vershynin, 2013b). In their case, they choose two vectors in $\mathbb{R}^2$, namely $x_1 = [\ 1 \quad 0\ ]$ and $x_1 = [\ 1 \quad 0.5\ ]$.[4] Their choice for P is the Bernoulli distribution on $\mathbb{R}^2$, which is purely atomic and assigns a weight of 0.25 to the four points $(1, 1), (1, -1), (-1, 1), (-1, -1)$. Now let us plot the half-planes H_{x_1}, H_{x_2} and their symmetric difference, which is the shaded region shown in Figure 1. Because none of the four points $(1, 1), (1, -1), (-1, 1), (-1, -1)$ (shown as red circles) belongs to the symmetric difference, x_1 and x_2 are indistinguishable in OBCS under this probability measure. In (Plan and Vershynin, 2013a), the authors conclude that OBCS cannot always recover an unknown vector x, depending on what P is. Indeed, x_1 and x_2 would be indistinguishable under *any* probability measure that assigns a value of zero to $H_{x_1}\Delta H_{x_2}$. Therefore, one must be careful to draw the right conclusion: When subsequent theorems in this paper show that OBCS is possible under *all* probability

4. They append a whole lot of zero components which are neglected here.

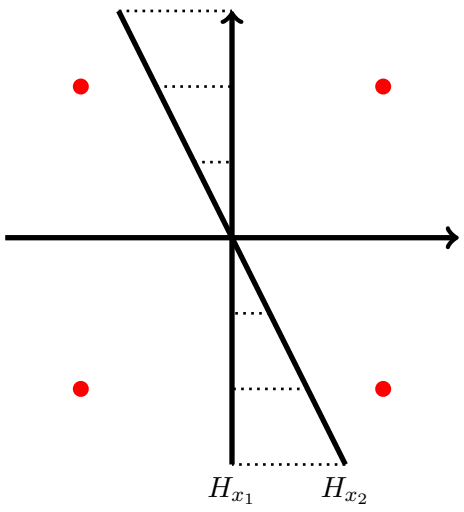

Figure 1: The half-planes H_{x_1}, H_{x_2}, and their symmetric difference.

measures on $\mathbb{R}^n$, *including* all purely atomic probability measures, what this means is that if $x \in \Sigma_k$, then OBCS will return a vector $\hat{x}$ such that $P(H_x \Delta H_{\hat{x}}) = 0$ whatever be P, and *not* that $x = \hat{x}$.

Now we examine the relationship of the generalization error $J(\hat{x}, x)$ to a couple of other quantities that are widely used in OBCS as error measures. First, define

$$\rho(\hat{x}, x) := E[|\text{sign}(\langle a, \hat{x}\rangle) - \text{sign}(\langle a, x\rangle)|, P], \tag{9}$$

where x is the true vector, $\hat{x}$ is its estimate. Note that $|\text{sign}(\langle a, \hat{x}\rangle) - \text{sign}(\langle a, x\rangle)|$ equals 0 or 2. Therefore we can also write

$$\rho(\hat{x}, x) = 2J(H_{\hat{x}}, H_x).$$

Next, we examine the relationship of $\rho(\hat{x}, x)$ to $\|\hat{x} - x\|_2$. Without loss of generality it can be assumed that both x and $\hat{x}$ have unit Euclidean norm. This can be achieved using some results from (Goemans and Williamson, 1995). Define

$$\alpha := \min_{\theta \in [0, 2\pi]} \frac{2}{\pi} \frac{\theta}{1 - \cos\theta} > 0.87856. \tag{10}$$

Then we have the following results.

Lemma 2 *Let P be any radially invariant probability measure on $\mathbb{R}^n$, and suppose $a \in \mathbb{R}^n$ is drawn at random according to P. Suppose $\|x\|_2 = \|\hat{x}\|_2 = 1$, and let $J(\hat{x}, x)$ denote the generalization error defined in* (5)*. Then*

$$\|\hat{x} - x\|_2^2 \leq \frac{4}{\alpha} J(\hat{x}, x). \tag{11}$$

Proof We make use of a couple of results from (Goemans and Williamson, 1995). First, (Goemans and Williamson, 1995, Lemma 3.2) states that

$$\Pr\{\text{sign}(\langle a, \hat{x}\rangle) \neq \text{sign}(\langle a, x\rangle)\} = \frac{1}{\pi}\arccos(\hat{x}^\top x),$$

Note that

$$J(\hat{x}, x) = \Pr\{\text{sign}(\langle a, \hat{x}\rangle) \neq \text{sign}(\langle a, x\rangle)\}.$$

Therefore the above is equivalent to

$$J(\hat{x}, x) = \frac{1}{\pi}\arccos(\hat{x}^\top x).$$

Second, (Goemans and Williamson, 1995, Lemma 3.4) states that

$$\frac{1}{\pi}\arccos(\hat{x}^\top x) \geq \frac{\alpha}{2}(1 - \hat{x}^\top x),$$

or equivalently

$$1 - \hat{x}^\top x \leq \frac{2}{\alpha}\arccos(\hat{x}^\top x).$$

Now note that, when both $\hat{x}$ and x are unit vectors, we have

$$\|\hat{x} - x\|_2^2 = \|\hat{x}\|_2^2 + \|x\|_2^2 - 2\hat{x}^\top x = 2(1 - \hat{x}^\top x).$$

Therefore

$$\begin{aligned} \|\hat{x} - x\|_2^2 &= 2(1 - \hat{x}^\top x) \\ &\leq \frac{4}{\alpha}\arccos(\hat{x}^\top x) = \frac{4}{\alpha}J(\hat{x}, x). \end{aligned}$$

■

3.4. PAC Learning via the Vapnik-Chervonenkis (VC) Dimension

One of the most useful concepts in PAC learning theory is defined next.

Definition 3 *A set $S \subseteq X$ of finite cardinality is said to be* **shattered** *by the concept class $\mathcal{C}$ if, for every subset $B \subseteq S$, there exists a concept $A \in \mathcal{C}$ such that $S \cap A = B$. The* **Vapnik-Chervonenkis- or VC-dimension** *of the concept class $\mathcal{C}$ is the largest integer d such that there exists a set S of cardinality d that is shattered by $\mathcal{C}$.*

Therefore a concept class $\mathcal{C}$ has VC-dimension d if two statements hold: (i) There exists a set of cardinality d that is shattered by $\mathcal{C}$, and *no set* of cardinality larger than d is shattered by $\mathcal{C}$. If there exist sets of arbitarily large cardinality that are shattered by $\mathcal{C}$, then its VC-dimension is defined to be infinite.

If a concept class has finite VC-dimension, then it is PAC learnable under *every probability distribution* on X. An algorithm is said to be **consistent** if it always produces a

hypothesis that classifies all the training samples correctly. In other words, an algorithm is consistent if the hypothesis G_m produced by applying the algorithm to the sequence $\{(c_i, I_T(c_i))\}_{i \geq 1}$ has the property that $I_T(c_i) = I_{G_m}(c_i)$ for all i and m. Note that a consistent algorithm always exists if the axiom of choice is assumed. However, in some situations, it is NP-hard or NP-complete to find a consistent algorithm.

With these notions in place, we have the following very fundamental result.

Theorem 4 *((Blumer et al., 1989); see also (Vidyasagar, 2003, Theorem 7.6)) Suppose a concept class $\mathcal{C}$ has finite VC-dimension. Then $\mathcal{C}$ is PAC learnable for every probability measure on X. Suppose that d is an upper bound for VC-dim($\mathcal{C}$), and let $\{A_m\}$ be any consistent algorithm. Moreover, the learning rate is bounded by*

$$r(m, \epsilon) \leq 2 \left(\frac{2em}{d} \right)^d 2^{-m\epsilon/2},$$

where e denotes the base of the natural logarithm. Therefore $r(m, \epsilon) \leq \delta$ if

$$m \geq \max \left\{ \frac{8d}{\epsilon} \lg \frac{8e}{\epsilon}, \frac{4}{\epsilon} \lg \frac{2}{\delta} \right\} \tag{12}$$

samples are chosen.

Note that the number of samples required to achieve an accuracy of ϵ with confidence $1 - \delta$ is $O((1/\epsilon) \ln(1/\delta))$. However, the main challenge in applying this result is in finding a consistent algorithm.

Theorem 4 shows that the finiteness of the VC-dimension of a concept class is a *sufficient* condition for PAC learnability. The next result shows that the condition is also necessary.

Theorem 5 *((Blumer et al., 1989; Ehrenfeucht et al., 1989); see also (Vidyasagar, 2003, Theorem 7.7)) Suppose a concept class $\mathcal{C}$ has VC-dimension $d \geq 2$. Then there exist probability measures on X such that any algorithm requires at least*

$$m \geq \left\{ \frac{d-1}{32\epsilon}, \frac{1-\epsilon}{\epsilon} \ln \frac{1}{\delta} \right\} \tag{13}$$

samples, in order to learn to accuracy ϵ and confidence δ.

4. Estimates of the VC-Dimension

This section contains the main results of the paper, namely, to obtain explicit estimates of the VC-dimension of half-spaces generated by k-sparse vectors.

Theorem 6 *Let $\mathcal{H}_k^n$ denote the set of half-spaces $H_k^n(x)$ in $\mathbb{R}^n$ generated by k-sparse vectors, as defined in (7). Then*

$$k(\lfloor \lg(n/k) \rfloor + 1) \leq \textit{VC-dim}(\mathcal{H}_k^n) \leq \lfloor 2k \lg(ne) \rfloor. \tag{14}$$

Remark: In (Neylon, 2006, Theorem 20), it is shown that if $k < 0.45\sqrt{n}$, then

$$\text{VC-dim}(\mathcal{H}_k^n) < 2k \lg n.$$

The right side of (14) can be rewritten as

$$\text{VC-dim}(\mathcal{H}_k^n) \leq 2k(\lg n + \lg e) \approx 2k(\lg n + 1.4).$$

Clearly, as n becomes larger, the additional term becomes insignificant. Moreover, there is no restriction on the magnitude of k in Theorem 6, as there is in (Neylon, 2006, Theorem 20). Also, the present Theorem 6 contains a lower bound on the VC-dimension in addition to an upper bound, which is not previously available in the literature.

Proof We begin with the upper bound in (14). It is shown that if a set $U = \{u_1, \ldots, u_l\} \subseteq \mathbb{R}^n$ is shattered by the collection of half-spaces $\mathcal{H}_k^n$, then $l \leq \lfloor 2k \lg(en) \rfloor$. The proof combines a few ideas that are standard in PAC learning theory, which are stated next.

The first result needed is (Dudley, 1978, Theorem 7.2).[5] It states the following: Suppose that $\mathcal{F}$ is a collection of functions mapping a given set Z into $\mathbb{R}$, such that $\mathcal{F}$ is a k-dimensional real vector space over the field $\mathbb{R}$. Define the associated collection of subsets of X by

$$\text{Pos}(f) := \{z \in Z : f(z) \geq 0\}, \text{Pos}(\mathcal{F}) := \{\text{Pos}(f), f \in \mathcal{F}\}.$$

Then VC-dim(Pos($\mathcal{F}$)) $= k$. To apply the above theorem to this particular instance, we fix the integer k as well as a support set $S \subseteq \{1, \ldots, n\}$ such that $|S| = k$, and choose $\mathcal{F}$ to be the set of functions $\{f(z) = \langle z, x \rangle : \text{supp}(x) \subseteq S\}$. This family of functions is clearly a k-dimensional linear space because the adjustable parameter here is the k-sparse vector x with support in the fixed set S. Therefore it follows that, if we define $\mathcal{H}_S = \{H_k^n(x) : \text{supp}(x) \subseteq S\}$, then VC-dim$(\mathcal{H}_S) = k$.

The next result needed is Sauer's lemma (Sauer, 1972), which states the following: Suppose $\mathcal{C}$ is a collection of subsets of X with finite VC-dimension d, and that $U = \{u_1, \ldots, u_l\} \subseteq X$ with $l > d$. Let $\mathcal{C} \cap U$ denote the collection $\{A \cap B : A \in \mathcal{C}, B \subseteq U\}$. Then

$$|\mathcal{C} \cap U| \leq \sum_{i=0}^{d} \begin{pmatrix} l \\ i \end{pmatrix} \leq \left(\frac{el}{d}\right)^d,$$

where e denotes the base of the natural logarithm. Strictly speaking, Sauer's lemma is the first inequality, which states that the number of subsets of U that can be generated by taking intersections with sets in the collection $\mathcal{C}$ is bounded by the summation shown. The second bound is derived in (Blumer et al., 1989). By applying Sauer's lemma to the problem at hand, it can be seen that, for a fixed support set S, the number of subsets of U that can be generated by intersecting with the collection of half-spaces $\mathcal{H}_S$ is bounded by $(el/k)^k$, because $\mathcal{H}_S$ has VC-dimension k. Note that a similar bound is derived in (Jacques et al., 2013), but without any reference to Sauer's lemma. Also, the result in (Jacques et al., 2013) is specifically for collections of half-spaces, whereas Sauer's lemma is applicable to completely general collections of sets.

5. Note that the definition of VC-dimension used in this reference is the VC-dimension as defined in Definition 3 *plus one*.

Now observe that $\mathcal{H}_k^n$ is just the union of the collections $\mathcal{H}_S$ as S ranges over all subsets of $\{1,\ldots,n\}$ with $|S| = k$. The number of such sets S is the combinatorial parameter n choose k, which is bounded by n^k.[6] Moreover, for each fixed support set S, the collection of subsets $\mathcal{H}_S \cap U$ has cardinality no larger than $(en/k)^k$, as shown above. Therefore

$$|\mathcal{H}_k^n \cap U| \leq \left(\begin{array}{c} n \\ k \end{array} \right) \left(\frac{el}{k} \right)^k \leq \left(\frac{nel}{k} \right)^k.$$

The final step in the proof comes from (Vidyasagar, 2003, Lemma 4.6), which states the following (see specifically item 2 of this lemma): Suppose $\alpha, \beta > 0$, $\alpha\beta > 4$ and $l \geq 1$. Then

$$l \leq \alpha \lg(\beta l) \implies l < 2\alpha \lg(\alpha\beta). \tag{15}$$

In the present instance, the collection of sets $\mathcal{H}_k^n \cap U$ has cardinality no larger $(nel/k)^k$, whereas U has 2^l subsets in all. Therefore, if U is shattered by the collection $\mathcal{H}_k^n$, then we must have

$$2^l \leq \left(\frac{nel}{k} \right)^k,$$

or, after taking binary logarithms,

$$l \leq k \lg \frac{nel}{k},$$

which is of the form (15) with $\alpha = k, \beta = ne/k$. Substituting these values into (15) leads to the conclusion that

$$l \leq 2k \lg(ne),$$

provided $\alpha\beta = ne \geq 4$, which holds if $n \geq 2$. Because l is an integer, we can replace the right side by its integer part, which leads to the upper bound in (14).

Now we turn our attention to the lower bound. First we consider the simple case where $k = 1$. Given n, define $l = \lfloor \lg n \rfloor + 1$, so that $n \geq 2^{l-1}$. The first step is to show that the set of half-spaces $\mathcal{H}_1^n$ generated by "one-sparse vectors" has VC-dimension l. Let $s = l - 1 = \lfloor \lg n \rfloor$, and enumerate the 2^s bipolar row vectors in $\{-1, 1\}^s$ in some order, call them $v_1, \ldots v_{2^s}$. Now define the $n \times l$ matrix

$$M = \left[\begin{array}{c|c} 1 & v_1 \\ \vdots & \vdots \\ 1 & v_{2^s} \\ 0_{(n-2^s)\times 1} & 0_{(n-2^s)\times s} \end{array} \right] \in \mathbb{R}^{n \times l}. \tag{16}$$

In other words, the matrix M has a first column of ones, and then the 2^s bipolar vectors in $\{-1, 1\}^s$ in some order, padded by a block of zeros in case $n > 2^{l-1}$. As we shall see below, the "padding" is not used. Now define $U = \{u_0, u_1, \ldots, u_s\}$ denote the $s + 1 = l$ columns of the matrix M. Note that for notational convenience we start numbering the columns with 0 rather than 1. It is claimed that the collection of half-spaces $\mathcal{H}_1^n$ shatters this set U, thus showing that VC-dim$(\mathcal{H}_1^n) \geq l$.

6. Actually n^k is a pretty crude estimate, but as we shall see, it is good enough.

To show that the set U is shattered, let $B \subseteq U$ be an arbitrary subset. Thus B consists of some columns of the matrix M. We examine two cases separately. First, suppose $u_0 \in B$. Then we associate a unique integer r between 1 and 2^s as follows. Define a bipolar vector $\mathbf{i}_B \in \{-1,1\}^s$ by $i_j = 1$ if $u_j \in B$, and $i_j = -1$ if $u_j \notin B$. This bipolar vector $\mathbf{i}_B$ must be one of the vectors $v_1, \ldots, v_{2^s}$. Let r be the unique integer such that $\mathbf{i}_B = v_r$. Define the vector $x \in \mathbb{R}^n$ such that $x_r = 1$, and the remaining elements of x are all zero, and note that $x \in \Sigma_1$. Then $\langle u_0, x\rangle = 1$, while $\langle u_j, x\rangle = 1$ if $u_j \in B$ and $\langle u_j, x\rangle = -1$ if $u_j \notin B$. Therefore the associated half-space $H(x)$ includes precisely the elements of the specified set B. Next, suppose $u_0 \notin B$; in this case we basically flip the signs. Thus the bipolar vector $\mathbf{i}_B \in \{-1,1\}^s$ is chosen such that $i_j = -1$ if $u_j \in B$, and $i_j = 1$ if $u_j \notin B$. If this bipolar vector corresponds to row r in the ordering of $\{-1,1\}^s$, we choose $x \in \Sigma_1$ to have a -1 in row r and zeros elsewhere. This argument shows that the set $\mathcal{H}_1^n$ generated by all one-sparse vectors x has VC-dimension of at least $\lfloor \lg n \rfloor + 1$, which is consistent with the left side of (14) when $k = 1$.

To extend the above argument to general values of k, suppose n and k are specified, and define $l = \lfloor \lg(n/k) \rfloor + 1$ and $s = l - 1 = \lfloor \lg(n/k) \rfloor$. Then $n/k \geq 2^s$, or equivalently, $n \geq k2^s$. Define matrices $M_1, \ldots, M_k \in \{-1,1\}^{2^s \times l}$ in analogy with (16). Then define a matrix $M \in \{0,1\}^{n \times kl}$ as a block-diagonal matrix containing $M_1, \ldots, M_k$ on the diagonal blocks, padded by an appropriate number of zero rows so that the number of rows equals n. In other words, M has the form

$$M = \begin{bmatrix} M_1 & 0_{2^s \times l} & \ldots & 0_{2^s \times l} \\ 0_{2^s \times l} & M_2 & \ldots & 0_{2^s \times l} \\ \vdots & \vdots & \vdots & \vdots \\ 0_{2^s \times l} & \ldots & 0_{2^s \times l} & M_k \\ & 0_{(n-k2^s) \times kl} & & \end{bmatrix} \in \mathbb{R}^{n \times kl}.$$

Define U to be the set of columns of the matrix M, and note that $|U| = kl = k(\lfloor \lg(n/k) \rfloor + 1)$. It is now shown that the set U is shattered by the collection $\mathcal{H}_k^n$ of half-spaces generated by k-sparse vectors. Partition U as $U_1 \cup U_2 \cup \ldots \cup U_k$, where each U_i consists of l column vectors. Then any specified subset $B \subseteq U$ can be expressed as a union $B_1 \cup \cdots \cup B_k$ where $B_i \subseteq U_i$ for each i. Now it is possible to mimic the arguments of the previous paragraph to show that the set U can be shattered by the collection of half-spaces $\mathcal{H}_k^n$. For each subset B_i, identify an integer r_i between 1 and 2^s such that the bipolar vector $\mathbf{i}_{B_i}$ is the r_i-th in the enumeration of $\{-1,1\}^s$. For each index i between 1 and k, let $x_i \in \mathbb{R}^{2^s}$ contain a 1 in row r_i and zeros elsewhere. Define $x \in \mathbb{R}^n$ by stacking x_1 through x_k, followed by $n - k2^s$ zeros. This shows that it is possible to shatter a set of cardinality $k(1 + \lfloor \lg(n/k) \rfloor)$, which is the right inequality in (14). ■

Theorem 6 is applicable to the case where measurements are of the form $\text{sign}(\langle a_i, x\rangle)$. Such measurements can at best lead to the recovery of the direction of a k-sparse vector x, but not its magnitude. In situations where it is desired to recover a sparse vector in its entirety, the measurements are changed to

$$y_i = \text{sign}(\langle a_i, x\rangle + b_i\theta), \tag{17}$$

where x varies over $\Sigma_k \subseteq \mathbb{R}^n$ and $\theta \in \mathbb{R}$. The concept class in this case is given by

$$H_k^{n+1}(x,\theta) = \{(a,b) \in \mathbb{R}^{n+1} : \langle a, x\rangle + b\theta \geq 0\}. \tag{18}$$

By inspecting the equation (17), we deduce that

$$\mathcal{H}_k^n(x) \subseteq \mathcal{H}_k^{n+1}(x) \subseteq \mathcal{H}_{k+1}^{n+1}([x\ \theta]^\top), \tag{19}$$

where $x \in \Sigma_k$ and $\theta \in \mathbb{R}$ so that the vector $[x\ \theta]^\top$ is $k+1$-sparse. We are now ready to state a bound on the VC-dimenstion of the collection of half-spaces $\mathcal{H}_k^{n+1}$.

Theorem 7 *Let $\mathcal{H}_k^{n+1}$ denote the set of half-spaces $H_k^n(x)$ in $\mathbb{R}^n$ as defined in* (17)*. Then*

$$k(1 + \lfloor \lg(n/k) \rfloor) \leq \textit{VC-dim}(\mathcal{H}_k^{n+1}) \leq \lfloor 2(k+1)\lg(e(n+1)) \rfloor. \tag{20}$$

Proof From (19) we conclude that

$$\text{VC-dim}(H_k^n) \leq \text{VC-dim}(\mathcal{H}_k^{n+1}) \leq \text{VC-dim} H_{k+1}^{n+1}([x\ \theta]^\top),$$

then the desired result follows Theorem 6. ■

5. OBCS with Noisy Measurements

In this section we study the one-bit compressed sensing problem when the information available to the learner is a randomly "flipped" version of the true output $\text{sign}(\langle a_i, x\rangle)$ or $\text{sign}(\langle a_i, x\rangle + b_i)$, where x is an unknown k-sparse vector. A study of the problem of PAC learning with noisy measurements was initiated in (Valiant, 1985), shortly after the publication of (Valiant, 1984). Over the years several different models of PAC learning with noisy labels have been studied. Rather than present an exhaustive listing of these, we focus on just a few papers that are most germane to the version of the problem studied here.

In the case of noise-free measurements, the results on learnability were stated in terms of a consistent algorithm, which always exists if one were to assume the axiom of choice. In contrast, in the case where the labels are noisy, it might not be possible to construct a hypothesis that is consistent with the data. Therefore the notion of consistency is replaced by the notion of minimizing empirical risk. Suppose we are given a labelled sample sequence $\{(c_i, y_i) \in X \times \{0,1\}\}_{i\geq 1}$. Suppose $F \in \mathcal{C}$ is a hypothesis. Then the **empirical risk** of the hypothesis with respect to this labelled sequence, after m samples, is defined as

$$\hat{J}_m(T,F) := \frac{1}{m}\sum_{i=1}^m |y_i - I_F(c_i)|. \tag{21}$$

Definition 8 *An algorithm $\{A_m\}_{m\geq 1}$is said to **minimize empirical risk**, or to be a **MER algorithm**, if for all sample sequences $\{(c_i, y_i) \in X \times \{0,1\}\}_{i\geq 1}$, and all integers m, it is the case that*

$$\hat{J}_m(G_m) = \min_{F\in\mathcal{C}} \hat{J}_m(F), \tag{22}$$

where

$$G_m = A_m((c_1, y_1), \ldots (c_m, y_m))$$

is the output of the algorithm after m samples, given the sequence $\{(c_i, y_i) \in X \times \{0,1\}\}_{i\geq 1}$.

Note that if the labels are noise-free, then $y_i = I_T(c_i)$, and

$$\hat{J}_m(T,F) := \frac{1}{m}\sum_{i=1}^{m} |I_T(c_i) - I_F(c_i)|$$

is the empirical estimate of the distance $D_P(T,F)$. In this case, a MER algorithm becomes a consistent algorithm.

We begin by summarizing some relevant results from (Angluin and Laird, 1988). In this paper, the probability of a 1 turning into 0 is assumed to be the same as the probability of a 0 turning into a 1; call it α. Clearly α must be < 0.5 in order for the problem to be learnable, because if the labels are flipped with probability 0.5, then the labels are pure noise and do not convey any information. In this paper, it is *not* assumed that the flipping probability α is known; rather, it is assumed that an upper bound $\bar{\alpha} < 0.5$ for α is known. Moreover, a procedure is given for generating such an upper bound. Suppose $\epsilon, \delta \in (0,1)$ are specified, where ϵ is the accuracy and δ is the confidence, and the objective is to choose the number of samples m to be sufficiently large that $r(m,\epsilon) \leq \delta$, where the learning rate $r(m,\epsilon)$ is defined in (6). The procedure given in this paper generates an upper bound $\bar{\alpha} < 0.5$ for α with confidence $\geq 1 - \delta/2$ using $r = 1 + \lceil \lg[(1-2\alpha)^{-1}] \rceil$ rounds. In case the concept class $\mathcal{C}$ has finite cardinality, say $|\mathcal{C}| \leq N$, then the following result can be established.

Theorem 9 *(See (Angluin and Laird, 1988, Theorem 2).) Given $\epsilon, \delta > 0$, draw*

$$m \geq \frac{2}{\epsilon^2(1-2\bar{\alpha})^2} \ln \frac{2N}{\delta} \tag{23}$$

i.i.d. samples according to an unknown probability measure $\mathcal{P} \in \mathcal{P}^*$. Let $T \in \mathcal{C}$ be any unknown target concept, and let G_m denote the output of a MER algorithm based on m noisy labels. Then

$$\sup_{P \in \mathcal{P}^*} \max_{T \in \mathcal{C}} P^m\{x \in X^m : d_P(G_m, T) \geq \epsilon\} \leq \delta. \tag{24}$$

If an infinite concept class $\mathcal{C}$ has finite VC-dimension, then for each $\epsilon > 0$ and $P \in \mathcal{P}^*$, $\mathcal{C}$ has a finite ϵ-cover with respect to the metric d_P, whose cardinality does not depend on $P \in \mathcal{P}^*$. Using this fact and Theorem 9, it is possible to prove bounds analogous to those in (25) for concept classes with finite VC-dimension. This is done in (Laird, 1988).

Next we discuss the paper (Natarajan et al., 2013). In this paper, the probability of a 1 becoming 0 is *not* assumed to equal the probability of a 0 becoming a 1. However, it is assumed that *both probabilities are known*. This allows the authors to construct an unbiased estimator for the true label, which may not be possible if only bounds for these probabilities are known.

Finally, we mention in passing the paper (Simon, 1996), even though the learning problem studied there is more general than the one studied here. Let $\mathcal{F} \subseteq [0,1]^X$ be a family of functions mapping X into $[0,1]$, and let $f \in \mathcal{F}$ be fixed but unknown. Let $P \in \mathcal{P}^*(X)$ be an unknown probability measure on X, and generate samples $x_1, \ldots, x_m$ in X that are i.i.d. according to P. For each sample x_j, a corresponding label $l_j \in \{0,1\}$ is generated according to

$$l_j = \begin{cases} 1, & \text{w.p. } f(x_j), \\ 0, & \text{w.p. } 1 - f(x_j). \end{cases}$$

From the data $\{(x_j, l_j)\}_{i=1}^m$, the objective is to cconstruct an approximation to $f(\cdot)$.

Now we come to the objective of the present section. By focusing on the *prediction error* (defined below) instead of the distance $d_P(G_m, T)$, it is possible to streamline the statement of the theorem.

To make the problem formulation precise, we use the notation in Section 3.1, whereby X is a set, $\mathcal{S}$ is a σ-algebra of subsets of X, and P is a probability measure on X. To incorporate the randomness, we enlarge X by defining $X_N = X \times \{0,1\}$ as the sample space; define $\mathcal{S}_N$ to be the σ-algebra of subsets in X_N generated by cylinder sets of the form $S \times \{0\}$ and $S \times \{1\}$ for all $S \in \mathcal{S}$; and define a probability measure P_N on X_N by defining

$$P_N(S \times \{0\}) = (1-\alpha)P(S), P_N(S \times \{1\}) = \alpha P(S). \tag{25}$$

Let (c, L) denote a typical element in the sample space X_N. Then

$$\Pr\{L = 0\} = P_N(X \times \{0\}) = 1 - \alpha, \Pr\{L = 1\} = P_N(X \times \{1\}) = \alpha.$$

Here the event $L = 0$ corresponds to the label not being flipped, while the event $L = 1$ corresponds to the label being flipped. It is clear that P_N is a product measure, so that the flipping of labels is independent of the generation of training samples.

Learning takes place as follows: Independent samples $\{(c_i, \omega_i) \in X \times \{0,1\}\}_{i \geq 1}$ are generated in accordance with the above probability measure P_N. Let T be a fixed but unknown target concept. Then for each i, a label y_i is generated as

$$y_i = |I_T(c_i) - \omega_i|. \tag{26}$$

This is equivalent to saying that $y_i = I_T(c_i)$ with probability $1 - \alpha$ and $y_i = 1 - I_T(c_i)$ with probability α. As before, an algorithm is an indexed family of maps $A_m : (X \times \{0,1\})^m \to \mathcal{C}$ for each $m \geq 1$. The algorithm A_m is applied to the set of labelled samples $\{(c_i, y_i)\}_{i=1}^m$, giving rise to a hypothesis G_m.

To assess how well a hypothesis F (however it is derived) approximates the unknown target concept T, we generate a random test input $x \in X$ according to P, and then predict that the oracle output on x will be $I_F(x)$. The error criterion therefore is the **prediction error**, and equals

$$J_N(T, F) := E[|f(I_T(x)) - I_F(x)|, P_N], \tag{27}$$

where $f(I_T(x))$ is the noisy label and $I_F(x)$ is the indicator function of F. The premise in the above definition is that, while the oracle output is noisy, our prediction is not noisy.

Note that, for a given $x \in X$, the quantity $|f(I_T(x)) - I_F(x)|$ equals $|I_T(x) - I_F(x)|$ with probability $1 - \alpha$, and equals $1 - |I_T(x) - I_F(x)|$ with probabilty α. Therefore

$$\begin{aligned} J_N(T, F) &= \int_X [(1-\alpha)|I_T(x) - I_F(x)| + \alpha(1 - |I_T(x) - I_F(x)|)]P(dx) \\ &= \alpha + (1 - 2\alpha) \int_X |I_T(x) - I_F(x)|P(dx) \\ &= \alpha + (1 - 2\alpha) d_P(T, F). \end{aligned} \tag{28}$$

Therefore, if we were to define the minimum achievable prediction error $J_N^*(T)$ as

$$J_N^*(T) = \min_{F \in \mathcal{C}} J_N(T, F),$$

then $J_N^*(T) = \alpha$ for each $T \in \mathcal{C}$. Therefore, to quantify the performance of a learning algorithm, we compare the actual prediction error $J_N(G_m, T)$ with $J_N^*(T)$, where G_m is the output of the algorithm based on m samples. With these observations, we can state the following result.

Theorem 10 *Suppose $\mathcal{C}$ is a concept class of subsets of X with VC-dim($\mathcal{C}$) $\leq d$, and let $T \in \mathcal{C}$ be a fixed but unknown target concept. Suppose $P \in \mathcal{P}^*$ is a fixed but unknown probability measure on X, $c_1, \ldots, c_m \in X$ are i.i.d. samples drawn according to P, and let G_m be the output of a MER algorithm. Then*

$$\Pr\{J_N(T, G_m) > J_N^*(T) + \epsilon\} \leq c(m, \epsilon), \tag{29}$$

where

$$c(m, \epsilon) = \left[4\left(\frac{0.2em}{d}\right)^{10d} + 1\right] \exp(-0.08m\epsilon^2). \tag{30}$$

Because the estimates derived above are broadly similar to those in (Angluin and Laird, 1988; Laird, 1988), we omit the proof of Theorem 10. The proof can be found in (Ahsen and Vidyasagar, 2017).

Note that the bound for the learning rate in Theorem 10 is qualitatively similar to that in Theorem 4. Therefore the bound (30) can be used to show the following: The quantity $\Pr\{J_N(T, G_m) > J_N^*(T) + \epsilon\}$ can be made smaller than a specified constant δ by choosing

$$m \geq \max\{O((1/\epsilon)\lg(1/\epsilon)), O((1/\epsilon)\lg(1/\delta))\} \tag{31}$$

samples.

Note that the error rate α does not appear explicitly in the right side of (30). However, if we examine (28), we see that α *does* appear in the estimate for the error $d_P(G_m, T)$. Specifically

$$d_P(G_m, T) = \frac{J_N(G_m, T) - J_N^*(T)}{1 - 2\alpha}.$$

Therefore

$$d_P(G_m, T) \leq \epsilon \iff J_N(G_m, T) - J_N^*(T) \leq (1 - 2\alpha)\epsilon. \tag{32}$$

Substituting this into Theorem 10 gives the following estimate.

Corollary 11 *Let all symbols be as in Theorem 10. Then*

$$\Pr\{d_P(G_m, T) \geq \epsilon\} \leq c(m, (1 - 2\alpha)\epsilon) = c(m, \bar{\epsilon}), \tag{33}$$

where $\bar{\epsilon} = (1 - 2\alpha)\epsilon)$, and $c(\cdot, \cdot)$ is defined in (30).

In view of Corollary 11, we infer that, in order to make the quantity $\Pr\{d_P(G_m, T) \geq \epsilon\}$ smaller than a specified constant δ, it is sufficient to choose

$$m \geq \max\{O((1/\bar{\epsilon})\lg(1/\bar{\epsilon})), O((1/\bar{\epsilon})\lg(1/\delta))\} \tag{34}$$

samples. Thus, for a fixed mislabelling probability α, the rate of growth of the sample complexity with respect to ϵ and δ is of the same order as in the noise-free case. However, as $\alpha \to 0.5^-$, the constant under the O symbol becomes larger and larger.

Note that Corollary 11 continues to hold if α is replaced by any upper bound for α.

6. Discussion

In this paper, the problem of one-bit compressed sensing (OBCS) has been formulated as a problem in probably approximately correct (PAC) learning theory. In particular, it has been shown that the VC-dimension of the set of half-spaces in $\mathbb{R}^n$ generated by k-sparse vectors is bounded by $O(k \lg n)$. Therefore, in principle at least, the OBCS problem can be solved using only $O(k \lg n)$ samples. This is possible in principle even when the measurements are corrupted by noise, except that as the mislabelling probability α approaches 0.5, the constant under the O symbol becomes larger and larger. However, in general, it is NP-hard to find a consistent algorithm when measurements are free from noise, and to find an algorithm that minimizes empirical risk when measurements are noisy.

One of the main advantages of formulating OBCS as a problem in PAC learning is that extending these results to the case where the samples $\{a_i\}$ (or $\{(a_i, b_i)\}$ as the case may be) are not i.i.d. essentially "comes for free." It is now known that, if a concept class has finite VC-dimension, then empirical means converge to their true values not only for i.i.d. samples $\{a_i\}$ (or $\{(a_i, b_i)\}$ as the case may be), but also when the samples come from a β-mixing stochastic process, e.g., from a Markov process. The convergence result is established in (Nobel and Dembo, 1993), and explicit rates of convergence are proved in (Karandikar and Vidyasagar, 2002). As it is fairly straight-forward to adapt the various theorems given here to the case of β-mixing processes using the above-mentioned results, the details are omitted.

Acknowledgement

The authors thank the reviewers for very helpful comments.

References

Mehmet Eren Ahsen and Mathukumalli Vidyasagar. An approach to one-bit compressed sensing based on probably approximately correct learning theory. arXiv:1710.07973va1, 2017.

A. Ai, A. Lapanowski, Y. Plan, and R. Vershynin. One-bit compressed sensing with non-Gaussian measurements. *Linear Algebra and Its Applications*, 441:222–239, 2014.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2: 343–370, 1988.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.

P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *Proceedings of the Conference on Information Sciences and Systems*, 2008.

Petros T. Boufounos. Greedy sparse signal reconstruction from sign measurements. In *Proceedings of the Asilomar Conference on Signals, Systems, and Computation*, 2009.

Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, December 2005.

Emmanuel J. Candès, J. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications in Pure and Applied Mathematics*, 59(8): 1207–1223, August 2006.

Mark A. Davenport, Marco F. Duarte, Yonina C. Eldar, and Gitta Kutyniok. Introduction to compressed sensing. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 1–68. Cambridge University Press, Cambridge, UK, 2012.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4): 1289–1306, April 2006a.

D. L. Donoho. For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution. *Communications in Pure and Applied Mathematics*, 59(6):797–829, 2006b.

R. M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, 6(6):899–929, 1978.

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989.

Michael Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer-Verlag, New York, 2010.

Yonina C. Eldar and Gitta Kutyniok, editors. *Compressed Sensing: Theory and Applications*. Cambridge University Press, Cambridge, UK, 2012.

Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer-Verlag, 2013.

M. X. Goemans and D. P. Williamson. Improved algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42:1115–1145, 1995.

Ankit Gupta, Robert Nowak, and Benjamin Recht. Sample complexity for 1-bit compressive sensing and sparse classification. In *Proceedings of the International Symposium on Information Theory*, 2010.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, Taylor & Francis Group, Boca Raton, FL, 2015.

Laurent Jacques, Jason N. Laska, Petros T. Boufounos, and Richard G. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, April 2013.

Rajeeva L. Karandikar and M. Vidyasagar. Rates of uniform convergence of empirical means with mixingprocesses. *Statistics & Probability Letters*, 58:297–307, 2002.

Karin Knudson, Rayan Saab, and Rachel Ward. One-bit compressive sensing with norm estimation. *IEEE Transactions on Information Theory*, 62(5):2748–2758, 2016.

Philip D. Laird. *Learning from Good and Bad Data*. Kluwer Academic Press, 1988.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234, 1995.

N. Natarajan, I. Dhillon, P. Ravikumar, and A. Tewari. Learning with noisy labels. *Neural Information Processig Systems*, 26, 2013.

Deanna Needell and Joel A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

S. Negabhan, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, December 2012.

Tyler Neylon. Sparse solutions for linear prediction problems. Ph.D. thesis, Department of Mathematics, New York University, 2006.

A. Nobel and A. Dembo. Rates of uniform convergence of empirical means with mixing-processes. *Statistics & Probability Letters*, 17:169–172, 1993.

Yaniv Plan and Ronan Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66:1275–1297, 2013a.

Yaniv Plan and Ronan Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, January 2013b.

Irina Rish and Genady Grabarnik. *Sparse Modeling: Theory, Algorithms, and Applications*. CRC Press, Taylor & Francis Group, Boca Raton, FL, 2015.

N. Sauer. On the densities of families of sets. *Journal of Combinatorial Theory, Series A*, 13:145–147, 1972.

Hans Ulrich Simon. General bounds on the number of examples needed for learning probabilistic concepts. *Journal of Computer and System Sciences*, 52(2):239–254, 1996.

L. G. Valiant. Learning conjuctions of disjunctions. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence*, pages 560–566, 1985.

Leslie G. Valiant. A theory of the learnable. *Journal of the ACM*, 29(11):1134–1142, 1984.

V. N. Vapnik. *Statistical Learning Theory*. John Wiley, New York, 1998.

M. Vidyasagar. *A Theory of Learning and Generalization*. Springer-Verlag, London, 1997.

M. Vidyasagar. *Learning and Generalization: With Applications to Neural Networks and Control Systems*. Springer-Verlag, London, 2003.