1. Introduction

Let \( C \subseteq \{0, 1\}^X \) be a class of functions from \( X \rightarrow \{0, 1\} \). We say that a pair \((Y, y)\) is a \( C \)-labelled sample if \( Y \subseteq X \) is a multiset and \( y = c|_Y \) for some \( c \in C \). The size of the labelled set is the size of \( Y \). For an integer \( k \), let

\[
L_C(k) = \{ (Y, y) : (Y, y) \text{ \( C \)-labelled and } |Y| \leq k \}.
\]

In this notation, \( L_C(\infty) \) is the set of all finite \( C \)-labelled samples.

**Definition 1.** A sample compression is a pair of maps \( \kappa, \rho \). The compression map

\[
\kappa : L_C(\infty) \rightarrow L_C(k) \times Q
\]

takes \((Y, y)\) to \(((Z, z), q)\) where \( Z \subseteq Y, |Z| \leq k, \) and \( y|_Z = z \). The reconstruction map

\[
\rho : L_C(k) \times Q \rightarrow \{0, 1\}^X
\]

is such that for all \((Y, y) \in L_C(\infty)\)

\[
\rho(\kappa(Y, y))|_Y = y.
\]

The size of the compression scheme is \( k + \log |Q| \).

**Example 1.** Let \( C \subseteq \{0, 1\}^\mathbb{R} \) be the set of indicator functions for closed intervals. Then for any \( C \)-labelled sample \((Y, y)\) let \( \kappa(Y, y) \) be given by \( Z = \min \{ x \in Y : y(x) = 1 \} \cup \max \{ x \in Y : y(x) = 1 \} \). Given such a set \( Z \), define

\[
f : Y \rightarrow \{0, 1\}
\]

by

\[
f(x) = \begin{cases} 
1 & \text{if } \min \{ z : z \in Z \} \leq x \leq \max \{ z : z \in Z \} \\
0 & \text{otherwise}
\end{cases}
\]

Then \( f|_Y = y \), and so this yields a compression scheme of size 2.

**Example 2.** Let \( C \subseteq \{0, 1\}^X \) be a class of functions lives in a vector space of rank \( r \) in \( \mathbb{R}^X \). That is, there exists \( r \) elements of \( C \) that span the entire class, and no such \( r - 1 \) elements. Then there is a size \( r \) compression scheme with no side information. Given any \( C \)-labelled sample \( Y, C|_Y \) has rank at most \( r \), and so let \( Z_Y \) be a set of columns of size \( r \) that span \( C|_Y \). Then we can uniquely determine \( c : Y \rightarrow \{0, 1\} \) given \( c|_{Z_Y} \). This is because if \( c_1, c_2 \) have the same restriction to \( Z_Y \), then since \( Z_Y \) spans the column space, the columns associated to \( c_1 \) and \( c_2 \) on \( Y \) must be identical.

Let’s recall the definition of VC-dimension and the fundamental theorem of statistical machine learning.
Definition 2. We say that \( Y \subseteq X \) is shattered by \( \mathcal{C} \) if for every \( f \in \{0, 1\}^Y \) there exists \( h \in \mathcal{C} \) such that \( h|_Y = f \), or in other words, if \( \mathcal{C}|_Y = \{0, 1\}^Y \). The VC-dimension (or Vapnik-Chervonenkis dimension) is the maximum size of a shattered subset of \( X \).

Theorem 1. (Fundamental theorem of machine learning) If \( \mathcal{C} \subset \{0, 1\}^X \) has VC-dimension \( d \), then \( \mathcal{C} \) is properly PAC-learnable with sample complexity
\[
m = O \left( \frac{d \log \left( \frac{2}{\epsilon} \right)}{\epsilon} + \frac{1}{\mu} \log \left( \frac{2}{\delta} \right) \right).\]
That is, there exists a learning map
\[
H : L_C(m) \to \{0, 1\}^X
\]
such that for every \( c \in \mathcal{C} \) and for every probability distribution \( \mu \) over \( X \)
\[
\mathbb{P}_{Y \sim \mu^m} [\mu(\{x \in X : h_Y(x) \neq c(x)\}) \leq \epsilon] \geq 1 - \delta
\]
where \( h_Y = H(Y, y) \).

If there exists a sample compression scheme for \( \mathcal{C} \) of size \( k \), then \( \mathcal{C} \) is PAC-learnable and has VC-dimension at most \( 8k \).

Theorem 2. (Littlestone-Warmuth 1986) Let \( \mathcal{C} \subset \{0, 1\}^X \), and let \( \kappa, \rho \) be a sample compression scheme for \( \mathcal{C} \) of size \( k \). Let
\[
m \geq \frac{8}{\epsilon} \left( k \log \left( \frac{2}{\epsilon} \right) + \log \left( \frac{1}{\delta} \right) \right)
\]
Then the learning map
\[
H : L_C(m) \to \{0, 1\}^X
\]
defined by \( H(Y, y) = \rho(\kappa(Y, y)) \) PAC-learns \( \mathcal{C} \) with \( m \) samples. That is, for every \( c \in \mathcal{C} \) and for every probability distribution \( \mu \) over \( X \)
\[
\mathbb{P}_{Y \sim \mu^m} [\mu(\{x \in X : h_Y(x) \neq c(x)\}) \leq \epsilon] \geq 1 - \delta
\]
where \( h_Y = H(Y, y) \).

Proof. We will prove that the VC-dimension is at most \( 8k \). Suppose that the VC-dimension of \( \mathcal{C} \) is \( d > 8k \). Then there exists \( Y \subseteq X \) of size \(|Y| = 8k\) such that \( \mathcal{C}|_Y \) yields all possible functions from \( Y \) to \( \{0, 1\} \). We will use a counting argument to show that for any compression scheme of size \( k \), there are distinct \( c_1, c_2 \in \mathcal{C} \) that cannot be distinguished, and hence that cannot be uncompressed. Given a compression mapping into \( L_C(l) \times Q \), there are at most
\[
|Q| \sum_{i=0}^{l} \binom{8k}{i} 2^i
\]
possible triples \(((Z, c|_Z), q)\) where \( Z \subset Y \) has size at most \( l \), \( c \) is some function in \( \mathcal{C} \) restricted to \( Z \), and \( q \in Q \) is some element of \( Q \). Since we have a sample compression scheme of size \( k \), we must have that \(|Q| \leq 2^{k-l} \), and so \( \kappa \) compresses the set of all functions on \( Y \), which has size \( 2^{8k} \), to a set of size at most
\[
2^{k-l} \sum_{i=0}^{l} \binom{8k}{i} 2^i < 2^{k+l} \binom{8k}{k},
\]
where the inequality follows since \( \binom{8k}{i} \) is monotonic in \( i \), and \( 1 + 2 + 2^2 + \cdots + 2^i < 2^{i+1} \). This quantity is strictly less than \( 2^{8k} \) for all \( k \geq 1 \), and so the proof is complete. \( \square \)
In their paper, Littlestone and Warmuth asked:

**Problem 1.** (Littlestone-Warmuth 1986) Are there concept classes of finite dimension for which there is no scheme with bounded kernel size and bounded additional information?

This was elegantly answered by Shay and Moran in 2015, and we will present their proof in the next section.

**Theorem 3.** (Moran-Yehudayoff 2015) Let \( C \subset \{0,1\}^X \) be a class of VC-dimension \( d \). Then there exists a sample compression scheme for \( C \) of size at most \( 2^{O(d)} \).

Littlestone and Warmuth conjecture that this bound could be improved further to \( O(d) \) on the size of the compression scheme.

**Conjecture 1.** (Littlestone-Warmuth 1986) Let \( C \subset \{0,1\}^X \) be a class with VC-dimension \( d \). Then there is a sample compression scheme for \( C \) of size at most \( O(d) \).

2. Proof of Shay-Moran

Throughout we let \( \epsilon = \frac{1}{3} \) and \( \delta = \frac{1}{3} \), as this choice of parameters will be sufficient for our purposes. If \( C \) has VC-dimension \( d \), then it follows from theorem 1 that there exists \( s = O(d) \) and a function \( H \) such that for every \( c \in C \) and for every distribution \( \mu \), there exists \( Z \subset \text{supp}(\mu) \) such that

\[
\mu(\{x \in X : h_Z(x) \neq c(x)\}) \leq \frac{1}{3},
\]

where \( h_Z = H(Z,c|Z) \) is the result of the learning algorithm. To create a sample compression scheme, we need to use our learning algorithm

\[
H : L_C(s) \to \{0,1\}^X
\]

where \( s = O(d) \). Given a \( C \)-labelled class \( (Y,y) \), consider the subset \( Z \subset Y, |Z| = s \) for which \( h_Z \) has the minimal error on \( Y \). We could hope to compress \( Y \to Z \), and then reconstruct \( y \) using our learning algorithm \( H \). However, since \( h_Z \) is not guaranteed to be 100% accurate on \( Y \), this will not work. Instead, we will look at multiple subsets \( Z_1, \ldots, Z_k \subset Y, |Z_i| = s \), and the resulting functions \( h_{Z_1}, \ldots, h_{Z_k} \), and ask them to vote on the value of \( y(x) \) for \( x \in Y \). In this case \( Z = \bigcup_{i=1}^k Z_i \) and our side information \( Q \) allows us to recover \( Z_i \) from \( Z \). Note in particular that there are many encoding schemes that allow us to take

\[
|Q| \leq (1 + sk)^{1+sk},
\]

where \( s = O(d) \), and so a bound on \( k \) is critical. To guarantee that the vote always returns the correct answer, we will use Von Neumann’s Min-Max Theorem.

**Theorem 4.** Let \( M \in \mathbb{R}^{m \times n} \) be a real matrix. Then

\[
\min_{p \in \Delta^m} \max_{q \in \Delta^n} p^t M q = \max_{q \in \Delta^n} \min_{p \in \Delta^m} p^t M q,
\]

where \( \Delta^\ell \) is the set on distributions on \( \{1, \ldots, \ell\} \).

**Corollary 1.** Suppose that for every \( p \in \Delta^m \), I can choose \( q \in \Delta^n \) such that

\[
p^t M q \geq c.
\]

Then there exists a distribution \( q^* \in \Delta^n \) such that

\[
p^t M q^* \geq c
\]
for any choice of \( p \).

Let \((Y, y)\) be any \( \mathcal{C} \)-labelled sample. By considering only distributions \( \mu \) which are supported on \( Y \), it follows that there exists \( Z \subset Y \) of size \( |Z| \leq s \) such that
\[
\mu(\{x \in Y : h_Z(x) = y(x)\}) > \frac{2}{3},
\]
where as before \( h_Z = H(Z, y|Z) \). Hence for any \( \mu \), there exists \( Z \) such that
\[
\mu(\{x \in Y : h_Z(x) = y(x)\}) > \frac{2}{3},
\]
and so it follows from the min-max theorem that there exists a distribution \( \nu \) over \( Z \subset Y \), \( |Z| = s \), such that for every \( x \in Y \)
\[
\nu(\{Z \subset Y : h_Z(x) = y(x)\}) > \frac{2}{3},
\]
This distribution \( \nu \) allows us to reconstruct \( y \) by looking at it on various subsets \( Z \subset Y \).

To finish our proof, we need an \( \epsilon \)-net result that allows us to approximate \( \nu \) as an average of only a handful of sets \( Z \).

**Theorem 5.** (Approximations for bounded VC-dimension) Let \( \mathcal{C} \subset \{0, 1\}^X \) be class of VC-dimension \( d \). Let \( \mu \) be a distribution on \( X \). Then for all \( \epsilon > 0 \), there exists a multiset \( W \subset X \) of size \( |W| \leq O\left(\frac{d}{\epsilon^2}\right) \) such that for all \( c \in \mathcal{C} \)
\[
\left| \mu(\{x \in X : c(x) = 1\}) - \frac{1}{|W|} |\{x \in W : c(x) = 1\}| \right| \leq \epsilon.
\]

**Definition 3.** Given \( \mathcal{C} \subset \{0, 1\}^X \), the dual class \( \mathcal{C}^* \) is defined as
\[
\mathcal{C}^* = \{c_x : x \in X\}
\]
where \( c_x : \mathcal{C} \to \{0, 1\} \) is the evaluation map \( c_x(c) = c(x) \).

**Theorem 6.** Let \( \mathcal{C} \subset \{0, 1\}^X \) be a class with dual VC-dimension \( d^* \). Let \( \nu \) be a distribution on \( \mathcal{C} \) and let \( \epsilon > 0 \). Then there exists a multiset \( F \subset \mathcal{C} \) of size
\[
|F| \leq O\left(\frac{d^*}{\epsilon^2}\right)
\]
such that for every \( x \in X \)
\[
\left| \nu(\{c \in \mathcal{C} : c(x) = 1\}) - \frac{1}{|F|} |\{f \in F : f(x) = 1\}| \right| \leq \epsilon.
\]

Applying this theorem with \( \epsilon = \frac{1}{8} \), it follows that there exists \( Z_1, \ldots, Z_k \subset Y \) such that for every \( x \in X \)
\[
\frac{1}{k} \left| \{i \in \{1, \ldots, k\} : h_{Z_i}(x) = y(x)\} \right| \geq \nu(\{Z \subset Y : h_Z(x) = y(x)\}) - \frac{1}{8} \\
\geq \frac{2}{3} - \frac{1}{8} \\
> \frac{1}{2}
\]
where \( k = O(d^*) \). To finish the proof, we use the following lemma:

**Lemma 1.** (Assouad 1983) Let \( \mathcal{C} \subset \{0, 1\}^X \) have VC-dimension \( d \). Then the dual class \( \mathcal{C}^* \) has VC-dimension \( d^* \) at most \( < 2^{d+1} \).
Proof. Suppose that $d^* \geq 2^{d+1}$. We will show that the VC-dimension of $\mathcal{C}$ is $\geq d + 1$. Then there is a subset $\mathcal{Y} \subset \mathcal{C}$ of functions of size $2^{d+1}$, and a subset $Y \subset X$ of points of size $2^{2d+1}$, such that the vectors

$$
\begin{bmatrix}
    c_1(x) \\
    \vdots \\
    c_{|Y|}(x)
\end{bmatrix}
$$

range over all possible $2^{2d+1}$ binary vectors as $x$ ranges over the points in $Y$. Looking at the $2^{d+1} \times 2^{2d+1}$ matrix $M$ whose rows are indexed by elements of $\mathcal{Y}$ and columns by elements of $Y$. Let $M'$ denote the $2^{d+1} \times (d + 1)$ matrix whose rows are the binary digits of the numbers $0, \ldots, 2^{d+1} - 1$ in order. For instance, when $d = 2$,

$$
M' = \begin{bmatrix}
    0 & 0 & 0 \\
    0 & 0 & 1 \\
    0 & 1 & 0 \\
    0 & 1 & 1 \\
    1 & 0 & 0 \\
    1 & 0 & 1 \\
    1 & 1 & 0 \\
    1 & 1 & 1
\end{bmatrix}.
$$

Since the columns of $M$ contain all possible binary vectors of length $2^{d+1}$, $M'$ must be a submatrix of $M$, and hence there exists a set of points of size $d + 1$ which is shattered by $\mathcal{C}$, and so the VC-dimension of $\mathcal{C}$ is $\geq d + 1$. \hfill \Box

Thus $k \leq 2^{d+1}$ and $\log Q \leq 2^{O(d)}$, and so the compression scheme has size at most $2^{O(d)}$, and we have proven the main theorem.