1 \( \epsilon \)-Net

Suppose \( X \) is a set with some distribution \( D \), and \( C \) is a class of boolean functions, each of which has the form \( F : X \to \{0, 1\} \). We can think of each function \( F \) as a concept, labeling each point in \( X \) as positive (1) or negative (0). The goal is to obtain a small subset \( S \subset X \) such that for each function \( F \in C \), if a large fraction (weighted according to distribution \( D \)) of points in \( X \) are marked as positive under \( F \), then there exists at least one point in \( S \) that is also marked positive under \( F \). We use \( E_x[F(x)] := E_{x \in D(X)}[F(x)] \) to denote the expectation of \( F(x) \), where \( x \) is a point drawn from \( X \) with distribution \( D \).

**Definition 1.1** An \( \epsilon \)-net \( S \) for a set \( X \) with distribution \( D \) under a class \( C \) of boolean functions on \( X \) is a subset satisfying the following:

For each \( F \in C \), if \( E_x[F(x)] \geq \epsilon \), then there exists \( x \in S \) such that \( F(x) = 1 \).

Trivially, we could take \( S := X \) as an \( \epsilon \)-net. However, we would want the cardinality of \( S \) to be small, even though \( X \) or \( C \) might be infinite.

We assume that we are able to sample points independently from \( X \) under distribution \( D \). The straightforward way to construct a net is to sample an enough number of points.

For \( 0 < \epsilon \leq 1 \), we define \( C_\epsilon := \{ F \in C : E_x[F(x)] \geq \epsilon \} \).

**Example**

Suppose \( X \) are points in the plane \( \mathbb{R}^2 \) with some distribution, and \( C \) is the class of functions, each of which corresponds to an axis-aligned rectangle that marks the points inside 1 and 0 otherwise. We would later see that for every \( 0 < \epsilon \leq 1 \), there is some finite sized \( \epsilon \)-net \( S_\epsilon \), i.e., if a rectangle contains more than \( \epsilon \) (weighted) fraction of points in \( X \), then it must contain a point in \( S_\epsilon \).

1.1 Simple Case: When \( C \) is finite

**Theorem 1.2** Suppose \( C \) is finite and \( S \) is a subset obtained by sampling from \( X \) independently \( m \) times. (There could be repeats, and so \( S \) could have size smaller than \( m \).) If \( m \geq \frac{1}{\epsilon} (\ln |C| + \ln \frac{1}{\delta}) \), then with probability at least \( 1 - \delta \), \( S \) is an \( \epsilon \)-net.

**Proof:** Observe that \( S \) is an \( \epsilon \)-net, if for all \( F \in C_\epsilon \), there is some point \( x \in S \) such that \( F(x) = 1 \). Fix any \( F \in C_\epsilon \), the probability that a point sampled from \( X \) would be labeled 1 is at least \( \epsilon \). Hence, the failure probability that all points in \( S \) are labeled 0 under \( F \) is at most \( (1 - \epsilon)^m \leq e^{-cm} \).
Using union bound, the probability that the set $S$ fails for some $F \in C_{\varepsilon}$ is at most $|C_{\varepsilon}| e^{-\varepsilon m} \leq |C| e^{-\varepsilon m}$, which is at most $\delta$, when $m \geq \frac{1}{\varepsilon} (\ln |C| + \ln \frac{1}{\delta})$.

1.2 Extending to Infinite $C$

Observe that for a fixed subset $S$ in $X$, if two functions $F$ and $F'$ agree on every point in $S$, then essentially they are the same from the viewpoint of $S$. Hence, for every fixed set $S$ of size $m$, there are effectively only $2^m$ boolean functions. However, there are still some issues.

1. There are still too many functions. Recall in the proof, we used the union bound to analyze the failure probability $|C| \cdot e^{-\varepsilon m} \leq 2^m \cdot e^{-\varepsilon m}$. However, this is not useful as the last quantity is larger than 1.

2. After we fix some $S$, there is no more randomness. Hence, we cannot even argue that the probability that $S$ is bad for even one $F$ is at most $(1 - \varepsilon)^m$.

For the first issue, we would add more assumptions to the class $C$ of functions to obtain a better guarantee. The second issue is technical and can be resolved by using the technique of conditional probability and expectation.

2 VC-Dimension: Limiting the Number of Boolean Functions on a Subset

Definition 2.1 Given a set $X$ and a class $C$ of boolean function on $X$, a subset $S \subseteq X$ is said to be shattered by $C$, if for all subsets $U$ of $S$, there exists $F \in C$ such that for all $x \in U$, $F(x) = 1$ and for all $x \in S \setminus U$, $F(x) = 0$.

The VC-dimension of $(X, C)$ is the maximum cardinality of a subset $S \subseteq X$ that is shattered by $C$. In other words, the VC-dimension of $(X, C)$ is at least $d$ if there exists $S \subseteq X$, where $|S| = d$, such that $S$ is shattered by $C$.

Example. Consider $X = \mathbb{R}^2$ and $C$ is the class where each function corresponds to an axis-aligned rectangle that labels each points inside it 1 and otherwise 0. Observer that $S = \{(1, 0), (-1, 0), (0, 1), (0, -1)\}$ can be shattered by $C$. However, one can show that no 5 points on the plane can be shattered by $C$.

Definition 2.2 Suppose $S \subseteq X$ and $F : X \to \{0, 1\}$. Then, the projection of $F$ on $S$ is the boolean function $F|_S : S \to \{0, 1\}$ such that for all $x \in S$, $F|_S (x) = F(x)$. Given a class $C$ of boolean functions, the projection $C(S)$ of $C$ on $S$ is the class $C(S) := \{ F|_S : F \in C \}$.

Given non-negative integers $m$ and $d$, we denote $\binom{m}{\leq d} := \sum_{i=0}^{d} \binom{m}{i}$.

Theorem 2.3 Suppose $C$ is a class of boolean functions on $X$ and the VC-dimension of $(X, C)$ is at most $d$. Let $S$ be a subset of $X$ of size $m$. Then, the cardinality of the projection $C(S)$ is at most $\binom{m}{\leq d}$. In particular, when $m \geq 2$ and $d \geq 2$, this is at most $m^d$.

Proof: We prove by induction on $d$ and $m$. For the base cases where $d$ and $m$ are small, we leave it to the readers to verify the claim. Suppose we have $S$, where $|S| = m > 1$, and the VC-dimension
of \((X, C)\) is \(d > 1\). We give an upper bound on \(|C(S)|\).

Let \(x \in S\) and define \(S' := S \setminus \{x\}\). Define \(C(S')^\dagger \subseteq C(S')\) to be the set of functions \(F\) in \(C(S')\) such that there exists \(F_1, F_2 \in C(S)\), where \(F_1\) and \(F_2\) disagree on \(x\) and \(F_1 |_{S'} = F_2 |_{S'} = F\).

Consider the projection of \(C\) on \(S'\). It follows that each function in \(C(S')^\dagger\) can be viewed as a “merge” of 2 functions in \(C(S')\). Hence, it follows that \(|C(S)| = |C(S')| + |C(S')^\dagger|\).

By induction hypothesis, we immediately have \(|C(S')| \leq \binom{m-1}{d-1}\).

We next show that the VC-dimension of \((S', C(S')^\dagger)\) is \(d - 1\). Suppose \(C(S')^\dagger\) shatters a subset \(U \subseteq S'\). Then, it follows immediately that \(C(S)\) shatters \(U \cup \{x\}\), which has size at most \(d\), since the VC-dimension of \((X, C)\) is at most \(d\). It follows \(|U| \leq d - 1\). Hence, by induction hypothesis \(|C(S')^\dagger| \leq \binom{m-1}{d-1}\).

By observing that \(\binom{m}{i} = \binom{m-1}{i} + \binom{m-1}{i-1}\), we conclude that \(|C(S)| \leq \binom{m}{d} + \binom{m-1}{d-1} = \binom{m}{d}\). \(\blacksquare\)

Here is the result relating VC-dimension of \((X, C)\) and the number of independent samples that is sufficient to form an \(\epsilon\)-net for \(X\) under \(C\).

**Theorem 2.4 (Number of Samples for Class with Bounded VC-Dimension)** Suppose \((X, C)\) has VC-dimension at most \(d\). Then, suppose \(S\) is a subset obtained by sampling from \(X\) independently \(m\) times (and removing repeated points). If \(m \geq \max\{\frac{1}{\delta} \log \frac{2}{\delta}, \frac{8d}{\epsilon} \log \frac{8d}{\epsilon}\}\), then with probability at least \(1 - \delta\), \(S\) is an \(\epsilon\)-net.

**Intuition.** Observe that \(|C(S)| \leq \binom{m}{d} \leq m^d\), for \(m \geq 2\) and \(d \geq 2\). Hence, if we use the “bogus” union bound, the failure probability would be at most \(|C(S)| \cdot e^{-cm} \leq m^d \cdot e^{-cm}\). When \(m\) is large enough as specified, this quantity is less than \(\delta\).

### 3 Conditional Probability and Expectation as Random Variables

We see that if \((X, C)\) has VC-dimension \(d\), then the projection of \(C\) on some subset \(S \subseteq X\) with \(|S| = m\) has size \(|C(S)| \leq m^d\). When we sample a subset \(S\), we would like to analyze the size of \(C(S)\), conditioned on the fact that \(S\) is sampled. We need some formal notation to analyze this.

**Definition 3.1 (Random Object)** Suppose \(\mathcal{P} = (\Omega, \mathcal{F}, Pr)\) is a probability space. A random object \(W\) taking values in some set \(\mathcal{U}\) is a function \(W : \Omega \rightarrow \mathcal{U}\). For \(u \in \mathcal{U}\), \(\{W = u\}\) is the event \(\{\omega \in \Omega : W(\omega) = u\}\).

**Example.**

1. A \(\{0,1\}\)-random variable is a special case when \(\mathcal{U} = \{0,1\}\).

2. Suppose we flip a fair coin repeatedly, and \(W\) is the outcome of the first 2 flips. In this case, \(\mathcal{U} = \{H,T\}^2\).

**Definition 3.2 (Conditional Probability as a Random Variable)** Suppose \(\mathcal{P} = (\Omega, \mathcal{F}, Pr)\) is a probability space, and \(A \in \mathcal{F}\) is an event. Let \(W : \Omega \rightarrow \mathcal{U}\) be a random object. Then, the conditional probability \(Pr[A|W]\) can be interpreted in two ways:

1. \(Pr[A|W] : \mathcal{U} \rightarrow [0,1]\) is a function such that for \(u \in \mathcal{U}\), \(Pr[A|W](u) := Pr[A|W = u]\).
2. \( \Pr[A|W] : \Omega \to [0,1] \) is a random variable defined by \( \Pr[A|W](\omega) := \Pr[A|W_\omega] \), where \( W_\omega := \{ \omega' \in \Omega : W(\omega') = W(\omega) \} \) is the event that \( W \) equals to \( W(\omega) \in \mathcal{U} \).

**Definition 3.3 (Conditional Expectation as a Random Variable)** Suppose \( \mathcal{P} = (\Omega, \mathcal{F}, \Pr) \) is a probability space, and \( Y : \Omega \to \mathbb{R} \) is a random variable. Let \( W : \Omega \to \mathcal{U} \) be a random object. Then, the conditional expectation \( E[Y|W] \) can be interpreted in two ways:

1. \( E[Y|W] : \mathcal{U} \to \mathbb{R} \) is a function such that for \( u \in \mathcal{U} \), \( E[Y|W](u) := E[Y|W = u] \).

2. \( E[Y|W] : \Omega \to \mathbb{R} \) is a random variable defined by \( E[Y|W](\omega) := E[Y|W_\omega] \), where \( W_\omega := \{ \omega' \in \Omega : W(\omega') = W(\omega) \} \) is the event that \( W \) equals to \( W(\omega) \in \mathcal{U} \).

Since the conditional probability \( \Pr[A|W] \) and the conditional expectation \( E[Y|W] \) are random variables themselves, we can take expectation of them.

**Fact 3.4** Let the event \( A \), the random variable \( Y \) and the random object \( W \) be defined as above. Then, \( E[\Pr[A|W]] = \Pr[A] \) and \( E[E[Y|W]] = E[Y] \).

**Example.** Consider the probability space associated with flipping a fair coin repeatedly. Let \( W \) be the outcome of the first 2 flips, and \( Y \) be the number of flips that a head first appears. As before, we have \( \mathcal{U} = \{H,T\}^2 \). Consider the conditional expectation \( E[Y|W] \).

1. We have \( E[Y|W = \{H,H\}] = 1 \), \( E[Y|W = \{H,T\}] = 1 \), \( E[Y|W = \{T,H\}] = 2 \). Finally, \( E[Y|\{T,T\}] = 2 \) and \( E[Y] = 4 \).

2. Hence, \( E[E[Y|W]] = \frac{1}{4}(1 + 1 + 2 + 4) = 2 = E[Y] \).

### 3.1 Using Conditional Probability to Bound Failure Probability

Recall that we are drawing independent samples from \( X \) to form a subset \( S \) of size \( m \) in the hope that \( S \) would be an \( \epsilon \)-net for the class \( C \) of functions. Suppose further that \( (X, C) \) has VC-dimension \( d \).

Let \( A \) be the event that \( S \) is not an \( \epsilon \)-net under \( C \). In particular, let \( A_F \) be the event that for all \( x \in S \), \( F(x) = 0 \). Recall that \( C_\epsilon := \{ C \in F : \epsilon \not\subseteq F \} \). We wish to find a good upperbound for \( \Pr[A] = \Pr[\cup_{F \in C_\epsilon} A_F] \).

Using conditional probability, we have \( \Pr[A] = E[\Pr[A|S]] \). Observe that if we fix \( S \), then the set \( S \) fails for the function \( F \in C \) if \( F \) fails for \( F' := F |_{S \in C(S)} \). Hence, \( \Pr[A|S] = \Pr[\cup_{F \in C_\epsilon} A_F|S] = \Pr[\cup_{F \in C_\epsilon(S)} A_{F'}|S] \leq \sum_{F' \in C_\epsilon(S)} \Pr[A_{F'}|S] \).

Observe that the summation contains at most \( |C_\epsilon(S)| \) at most \( |C(S)| \leq m^d \) terms. Hence, it suffices to give a good upperbound on \( p^* := \max_{F' \in C_\epsilon(S)} \Pr[A_{F'}|S] \). However, as we mention before, if we condition on \( S \), there is no more randomness, since \( \Pr[A_F|S] \) is either 0 or 1. Hence, we can have \( p^* = 1 \). We shall see next time how we can resolve this by introducing extra randomness in the analysis.
4 Homework Preview

1. VC-dimension of Axis-aligned rectangles.
   (a) Prove that no 5 points on the plane \( \mathbb{R}^2 \) can be shattered by the class \( C \) of axis-aligned rectangles (that map points inside a rectangle 1 and otherwise 0).
   (b) Compute the VC-dimension of the class \( C_k \) of \( k \)-dimensional axis-aligned rectangles in \( \mathbb{R}^k \). In particular, you need to find a number \( d \) such that there exist \( d \) points in \( \mathbb{R}^k \) that can be shattered by the \( C_k \), and prove that any \( d + 1 \) points in \( \mathbb{R}^k \) cannot be shattered by \( C_k \).

2. Conditional Expectation. Suppose \( Y : \Omega \rightarrow \mathbb{R} \) is a random variable and \( W : \Omega \rightarrow U \) is a random object defined on the same probability space \( (\Omega, \mathcal{F}, Pr) \). Prove that \( E[Y] = E[E[Y|W]] \). You may assume that both \( \Omega \) and \( U \) are finite.

3. Using \( \epsilon \)-Net for Learning. Suppose \( X \) is a set with some underlying distribution \( D \) and \( C \) is a class of boolean functions on \( X \), and the VC-dimension of \( (X, C) \) is \( d \). Moreover, suppose there is some function \( F_0 \in C \) that corresponds to some classifier that we wish to learn. The model we have is that we can sample a random \( x \in X \) and ask for the value \( F_0(x) \). After seeing \( m \) such samples \( S \) in \( X \), we pick a function \( F_1 \in C \) that agrees with \( F_0 \) on \( S \). The hope is that \( F_1 \) and \( F_0 \) would agree on most points in \( X \) (according to distribution \( D \)).
   (a) Define another class \( C' \) of boolean functions on \( X \) such that if \( S \) is an \( \epsilon \)-net under \( C' \), and \( F \in C \) is a function that disagrees with \( F_0 \) on more than \( \epsilon \) fraction (weighted according to \( D \)) of points in \( X \), then there exists some \( x \in S \) such that \( F(x) \neq F_0(x) \). Prove the VC-dimension of \( (X, C') \) for the class \( C' \) that you have constructed.
   (b) How many samples are enough such that with probability at least \( 1 - \delta \) the function \( F_1 \) returned disagrees with \( F_0 \) on at most \( \epsilon \) weighted fraction of points in \( X \)?