7.2 Epsilon-Approximations via Chaining

Our main theorem will be the following.

Theorem 7.3. *Let $(X, \mathcal{F})$ be a set system with $\text{VC-dim}(\mathcal{F}) \le d$, and $\epsilon > 0$ a given parameter. Then a uniform random sample A of X of size $O\left(\frac{d}{\epsilon^2}\right)$ is an ϵ-approximation for $\mathcal{F}$, with constant probability.*

Throughout the rest of this chapter, we will fix a set system $(X, \mathcal{F})$ with $|X| = n$ and $\text{VC-dim}(\mathcal{F}) \le d$, and a parameter $\epsilon > 0$.

Recall ϵ-approximations:

Definition 7.1. *Given a set system $(X, \mathcal{F})$, an ϵ-approximation is a subset $A \subseteq X$ such that for any set $S \in \mathcal{F}$, we have*

$$\left| \frac{|S|}{|X|} - \frac{|S \cap A|}{|A|} \right| \le \epsilon.$$

For the moment we treat the size of A as a parameter t, whose value will then be set during the proof as needed. There are three additional ideas in the proof, which we outline before we present the complete proof.

1. Getting independence from n.

To bound the failure probability of A, we will bound the probability of A being a failure for a fixed set $S \in \mathcal{F}$, and then use the union bound over all sets of $\mathcal{F}$. That brings in the dependence on $\log |\mathcal{F}| = O(d \log n)$. An easy 'pre-processing' step allows us to replace n by a function of $\frac{1}{\epsilon}$: we will first take a larger random sample A' of size $O\left(\frac{d}{\epsilon^2} \log \frac{1}{\epsilon}\right)$ uniformly from $(X, \mathcal{F})$. As seen before, with constant probability, A' is a $\frac{\epsilon}{2}$-approximation for $(X, \mathcal{F})$.

The next lemma shows that if we now compute an $\frac{\epsilon}{2}$-approximation for $(A', \mathcal{F}|_{A'})$, it will be an ϵ-approximation for $(X, \mathcal{F})$.

Lemma 7.2. *Given a set system $(X, \mathcal{F})$, let $A \subseteq X$ be an ϵ_1-approximation for $\mathcal{F}$. Further let $B \subseteq A$ be an ϵ_2-approximation for $(A, \mathcal{F}|_A)$. Then B is a $(\epsilon_1 + \epsilon_2)$-approximation for $(X, \mathcal{F})$.*

Proof. Fix any $S \in \mathcal{F}$. Then

$$\begin{aligned} A \text{ is an } \epsilon_1\text{-approximation for } (X, \mathcal{F}) \quad &\Longrightarrow \quad |A \cap S| &&= \frac{|S| \cdot |A|}{|X|} &&\pm \epsilon_1 \cdot |A|. \\ B \text{ is an } \epsilon_2\text{-approximation for } (A, \mathcal{F}|_A) \quad &\Longrightarrow \quad |B \cap (S \cap A)| &&= \frac{|S \cap A| \cdot |B|}{|A|} &&\pm \epsilon_2 \cdot |B|. \end{aligned}$$

As $B \subseteq A$, we have $|B \cap (S \cap A)| = |B \cap S|$, and so

$$|B \cap S| = |B \cap (S \cap A)| = \frac{\left(\frac{|S| \cdot |A|}{|X|} \pm \epsilon_1 \cdot |A|\right) \cdot |B|}{|A|} \pm \ \epsilon_2 \cdot |B|.$$

$$= \frac{|S| \cdot |B|}{|X|} \pm \ (\epsilon_1 + \epsilon_2) \cdot |B|.$$

□

Thus, we can take a sample A' and then work with $(A', \mathcal{F}|_{A'})$ instead of $(X, \mathcal{F})$. This allows us to assume that $|X| = O\left(\frac{d}{\epsilon^2} \log \frac{1}{\epsilon}\right)$, which implies that $|\mathcal{F}| = O\left(\left(\frac{d}{\epsilon^2} \log \frac{1}{\epsilon}\right)^d\right)$.

Next, we get rid of the 'log' term entirely.

2. Taking care of small-sized sets of $\mathcal{F}$.

From the definition of ϵ-approximations, for each $S \in \mathcal{F}$, we want

$$|A \cap S| = \frac{|S| \cdot t}{n} \pm \epsilon t.$$

As $\mathrm{E}\left[|A \cap S|\right] = \frac{|S| \cdot t}{n}$, A fails for S if the random variable $|A \cap S|$ is greater than ϵt away from its expectation.

We first recall Chernoff's bound:

Theorem (Chernoff's bound). *Given n independent random variables $X_i \in [0,1]$, $X = \sum_i X_i$,*

$$\delta \geq 0 : \qquad \Pr\left[X \geq (1+\delta) \cdot \mathrm{E}\left[X\right]\right] \leq e^{-\frac{\delta^2 \, \mathrm{E}[X]}{2+\delta}}$$

$$0 \leq \delta \leq 1 : \qquad \Pr\left[X \leq (1-\delta) \cdot \mathrm{E}[X]\right] \leq e^{-\frac{\delta^2 \, \mathrm{E}[X]}{2}}$$

In our case, we want

$$|A \cap S| = \frac{|S| \cdot t}{n} \pm \epsilon t = \frac{|S| \cdot t}{n} \left(1 \pm \frac{\epsilon n}{|S|}\right).$$

Thus $\delta = \frac{\epsilon n}{|S|}$. We now calculate the probability that $|A \cap S|$ is greater than $\left(1 + \frac{\epsilon n}{|S|}\right) \mathrm{E}\left[|A \cap S|\right]$.

Consider the case when δ is large, say $\delta = \frac{\epsilon n}{|S|} \geq 1$. Then $\frac{\delta^2}{2+\delta} = c'\delta$ for an absolute constant c', and so

$$\Pr\left[|A \cap S| \geq \left(1 + \frac{\epsilon n}{|S|}\right) \mathrm{E}\left[|A \cap S|\right]\right] \leq e^{-c'\left(\frac{\epsilon n}{|S|}\right) \cdot \frac{|S| t}{n}} = e^{-c' \epsilon t}. \tag{7.3}$$

As the number of sets in $\mathcal{F}$ are $\tilde{O}\left(\frac{1}{\epsilon^d}\right)$, we can set $t = \tilde{O}\left(\frac{1}{\epsilon}\right)$, and conclude that, with high probability, A fails for no set of $\mathcal{F}$.

3. Packing lemma takes care of large sets.

It remains to deal with the case of sets S with $|S| \geq \epsilon n$. This poses a problem, as the error interval—ϵt—is *independent* of the size of S. As the tail bounds will give the failure probability as a function of the size of S, for $|S|$ large enough, ϵt could occur inside this 'margin of error', and so is likely to happen.

To be specific, when $\delta = \frac{\epsilon n}{|S|} \leq 1$, we have $\frac{\delta^2}{2+\delta} \geq \frac{\delta^2}{3}$, and so

$$\Pr\left[|A \cap S| \geq \left(1 + \frac{\epsilon n}{|S|}\right) \mathrm{E}\left[|A \cap S|\right]\right] \leq e^{-\frac{1}{3}\left(\frac{\epsilon n}{|S|}\right)^2 \cdot \frac{|S|t}{n}} = e^{-\frac{1}{3}\frac{\epsilon^2 nt}{|S|}} \leq e^{-\frac{1}{3}\epsilon^2 t}.$$

Here we are forced to set $t = \Theta\left(\frac{1}{\epsilon^2}\log\frac{1}{\epsilon}\right)$ to take care of all the $\tilde{O}\left(\frac{1}{\epsilon^d}\right)$ sets in $\mathcal{F}$, and we're back to our old bound.

However, notice that in the last inequality above, we upper-bounded $|S|$ by n. *If* it were true that the number of sets in $\mathcal{F}$ of size $\Omega\left(n\right)$ were $O\left(1\right)$, then we'd still be fine, as then we could set $t = \Theta\left(\frac{d}{\epsilon^2}\right)$ as we only have to union over $O(1)$ such events. Unfortunately that is too optimistic[†].

However, in this worst-case scenario—say when considering sets of size $\Omega\left(n\right)$ in $\mathcal{F}$—it is clear that these sets have a lot of elements of X in common, and so we really need to union over a much smaller set of events. Thus we consider the sets of $\mathcal{F}$ by their sizes, and use the packing lemma to capture the fact that there are few sets that are the 'basic', and that most of the large sets are combinations of these basic sets.

However, the number of these basic sets increase with decreasing set sizes when using the packing lemma, and thus the bounds degrade as the set size decrease. Thus we will switch over to the previous analysis when the remaining sets have a small-enough size. This trade-off will give us our desired bound.

We now give the complete formal proof incorporating these ideas.

⋆ ⋆ ⋆

We will construct a series of maximal packings. For a parameter k to be fixed later, for each $i = 1, \ldots, k+1$, define

$$\mathcal{P}_i: \text{ a maximal } \left(\frac{n}{2^{2i}}\right)\text{-packing of } (X, \mathcal{F}).$$

[†] Consider when X is a set of n points in $\mathbb{R}^d$, and the subsets of $\mathcal{F}$ are induced by half-spaces. This has VC-dimension $d+1$. Further, there are $\binom{n}{d}$ distinct subsets produced by hyperplanes, one side of which has at least $\frac{n}{2}$ points. Thus, there are $\binom{n}{d} = \Omega\left(n^d\right)$ subsets of size at least $\frac{n}{2}$ in $\mathcal{F}$.

At the i-th level, by the packing lemma, we have

$$|\mathcal{P}_i| = O\left(\left(\frac{n}{n/2^{2i}}\right)^d\right) = O\left(2^{2di}\right).$$

For each $i \in [1,k]$ and $S \in \mathcal{P}_{i+1} \setminus \mathcal{P}_i$, by the maximality of $\mathcal{P}_i$, there exists a set $F_S \in P_i$ such that

$$|\Delta\left(S, F_S\right)| \leq \frac{n}{2^{2i}}.$$

For all $i = 1, \ldots, k$, define the sets

$$\mathcal{A}_i = \left\{S \setminus F_S \colon S \in \mathcal{P}_{i+1} \setminus \mathcal{P}_i\right\},$$
$$\mathcal{B}_i = \left\{F_S \setminus S \colon S \in \mathcal{P}_{i+1} \setminus \mathcal{P}_i\right\}.$$

Then we have

$$|\mathcal{A}_i|, |\mathcal{B}_i| \leq |\mathcal{P}_{i+1}| = O\left(2^{2d(i+1)}\right), \qquad \text{and} \qquad \forall\, S' \in \mathcal{A}_i \cup \mathcal{B}_i, \quad |S'| \leq \frac{n}{2^{2i}}.$$

Lemma 7.3. *For each $j \in [1,k]$ and $S \in \mathcal{P}_{j+1} \setminus \mathcal{P}_j$, there exist an initial set $I \in \mathcal{P}_1$ and sets*

$$A_1 \in \mathcal{A}_1,\ A_2 \in \mathcal{A}_2,\ \ldots, A_j \in \mathcal{A}_j \qquad \textit{and} \qquad B_1 \in \mathcal{B}_1,\ B_2 \in \mathcal{B}_2,\ \ldots,\ B_j \in \mathcal{B}_j$$

such that $A_i \cap B_i = \emptyset$ for all $i = 1, \ldots, j$, and

$$S = \Bigg(\cdots\Big(\big((I - B_1 + A_1) - B_2 + A_2\big) - B_3 + A_3\Big)\cdots - B_j + A_j\Bigg). \tag{7.4}$$

Furthermore, if A is an ϵ_0-approximation for all the sets of $\mathcal{P}_1$ and an ϵ_i-approximation for $\mathcal{A}_i \cup \mathcal{B}_i$, then A is an $(\epsilon_0 + 2\epsilon_1 + 2\epsilon_2 + \cdots + 2\epsilon_k)$-approximation for all the sets of $\mathcal{P}_k$.

Proof. As $S \in \mathcal{P}_{j+1} \setminus \mathcal{P}_j$, there exists $F_S \in \mathcal{P}_j$ is such that $|\Delta\left(S, F_S\right)| \leq \frac{n}{2^{2j}}$, and further we can write

$$S = F_S + (S \setminus F_S) - (F_S \setminus S) = F_S + A_j - B_j,$$

where $A_j \in \mathcal{A}_j$ and $B_j \in \mathcal{B}_j$.

We can again write $F_S \in \mathcal{P}_j$ as an addition/subtraction of sets in $\mathcal{A}_{j-1}$ and $\mathcal{B}_{j-1}$. Continuing like this recursively, we get that any set $S \in \mathcal{P}_{j+1} \setminus \mathcal{P}_j$ can be written as addition of j sets and subtraction of j sets, ending up at some set $I \in \mathcal{P}_1$. This gives Equality (7.4).

For the second part, assume that the set $A \subseteq X$ is an ϵ_0-approximation for $\mathcal{P}_1$ and an ϵ_i-approximation for the sets in $\mathcal{A}_i \cup \mathcal{B}_i$, for $i = 1, \ldots, k$. In other words,

$$\text{for all } S' \in \mathcal{A}_i, \mathcal{B}_i\colon \qquad |A \cap S'| = \frac{|S'| \cdot |A|}{n} \pm \epsilon_i \cdot |A|.$$

Assume that for any $S' \in \mathcal{P}_j \setminus \mathcal{P}_{j-1}$, we have

$$|A \cap S'| = \frac{|S'| \cdot |A|}{n} \pm \left(\epsilon_0 + 2 \sum_{i=1}^{j-1} \epsilon_i \right) \cdot |A|.$$

Then we have $|S| = |F_S| + |A_j| - |B_j|$, and so

$$\begin{aligned}
|A \cap S| &= |A \cap F_S| + |A \cap A_j| - |A \cap B_j| \\
&= \left(\frac{|F_S| \cdot |A|}{n} \pm \left(\epsilon_0 + 2 \sum_{i=1}^{j-1} \epsilon_i \right) \cdot |A| \right) + \left(\frac{|A_j| \cdot |A|}{n} \pm \epsilon_j \cdot |A| \right) - \left(\frac{|B_j| \cdot |A|}{n} \pm \epsilon_j \cdot |A| \right) \\
&= \left(\frac{(|F_S| + |A_j| - |B_j|) \cdot |A|}{n} \right) \pm \left(\epsilon_0 + 2 \sum_{i=1}^{j-1} \epsilon_i \right) \cdot |A| \pm \epsilon_j \cdot |A| \pm \epsilon_j \cdot |A| \\
&= \frac{|S| \cdot |A|}{n} \pm \left(\epsilon_0 + 2 \sum_{i=1}^{j} \epsilon_i \right) \cdot |A|.
\end{aligned}$$

□

We set $\epsilon_0 = \frac{\epsilon}{4}$ and for $i = 1, \ldots, k$,

$$\epsilon_i = \frac{\sqrt{i}}{4 \cdot 2^i} \cdot \epsilon.$$

We now calculate the probability that for any fixed index i, A is not an ϵ_i-approximation for some set in $\mathcal{A}_i \cup \mathcal{B}_i$. For a fixed set $S_i \in \mathcal{A}_i \cup \mathcal{B}_i$, this probability is

$$\Pr\left[|A \cap S_i| < \left(1 - \frac{\epsilon_i n}{|S_i|} \right) \mathrm{E}\left[|A \cap S_i| \right] \right] + \Pr\left[|A \cap S_i| > \left(1 + \frac{\epsilon_i n}{|S_i|} \right) \mathrm{E}\left[|A \cap S_i| \right] \right],$$

where $\mathrm{E}\left[|A \cap S_i|\right] = \frac{|S_i| t}{n}$. In our case, $\delta = \frac{\epsilon_i n}{|S_i|}$, and so Chernoff's bound gives

$$\Pr\left[|A \cap S_i| \le \left(1 - \frac{\epsilon_i n}{|S_i|} \right) \mathrm{E}\left[|A \cap S_i| \right] \right] \le \exp\left(-\frac{1}{2} \frac{\epsilon_i^2 n^2}{|S_i|^2} \frac{|S_i| t}{n} \right) = \exp\left(-\frac{1}{2} \frac{\epsilon_i^2 n t}{|S_i|} \right) \le e^{-\frac{1}{8} i \epsilon^2 t}.$$

At level i, there are $O\left(2^{2d(i+1)}\right)$ sets in $\mathcal{A}_i \cup \mathcal{B}_i$ over which we will use the union bound. As the above probability of failure of each fixed set of $\mathcal{A}_i \cup \mathcal{B}_i$ is at most $e^{-\Omega\left(i \cdot \epsilon^2 \cdot t\right)}$, it suffices to set $t = O\left(\frac{d}{\epsilon^2}\right)$, and we will end up with a geometric series, which when summed up over all levels, is less than 1.

However, the asymmetry of the Chernoff bound poses problems for the other direction:

$$\Pr\left[|A \cap S_i| \ge \left(1 + \frac{\epsilon_i n}{|S_i|} \right) \mathrm{E}\left[|A \cap S_i| \right] \right] \le \exp\left(- \frac{\frac{\epsilon_i^2 n^2}{|S_i|^2} \frac{|S_i| t}{n}}{2 + \frac{\epsilon_i n}{|S_i|}} \right)$$

$$= \exp\left(-\frac{\epsilon_i^2 nt}{2|S_i| + \epsilon_i n}\right) \leq \exp\left(-\frac{i\epsilon^2 nt}{2^{2i+4}\left(2 \cdot n/2^{2i} + \frac{\sqrt{i\epsilon n}}{2^{i+2}}\right)}\right)$$

$$= \exp\left(-\frac{i\epsilon^2 t}{2^5 + 4 \cdot 2^i\sqrt{i\epsilon}}\right).$$

The function $2^i\sqrt{i}\epsilon$ is increasing with the depth i of our decomposition scheme. We will fix k so that $2^k\sqrt{k}\epsilon \leq 1$, achieved by setting $k = \log\left(\frac{1}{\epsilon}\right)^{1/2}$. Then the probability that A fails to be an ϵ_i-approximation simultaneously for all the sets of $\mathcal{A}_i \cup \mathcal{B}_i$, for all $i = 1, \ldots, k$, can be upper-bounded by

$$\sum_{i=1}^{k} \sum_{S \in \mathcal{A}_i \cup \mathcal{B}_i} \Pr\left[A \text{ is not an } \epsilon_i\text{-approximation for } S\ \right]$$

$$\leq 2\sum_{i=1}^{k} |\mathcal{A}_i \cup \mathcal{B}_i| \cdot e^{-\frac{i\epsilon^2 t}{36}} \leq 2\sum_{i=1}^{k} O\left(2^{2d(i+1)}\right) \cdot e^{-2di} \ll 1,$$

where the second-last inequality follows by setting $t = \frac{72d}{\epsilon^2}$. Thus, with non-zero probability, A is an ϵ_i-approximation simultaneously for all the sets of $\mathcal{A}_i, \mathcal{B}_i$, for all $i = 1, \ldots, k$.

Now we finish up the proof by considering all possible $S \in \mathcal{F}$:

Case $S \in \mathcal{P}_1$**:** Noting that $|\mathcal{P}_1| = O\left(2^{2d}\right)$ and $\epsilon_0 = \frac{\epsilon}{4}$,

$$\Pr\left[A \text{ is not an } \epsilon_0\text{-approximation for } \mathcal{P}_1\ \right] \leq 2\sum_{S \in \mathcal{P}_1} e^{-\frac{1}{3}\frac{\epsilon_0^2 n^2}{|S|^2}\frac{|S|t}{n}}$$

$$= 2\sum_{S \in \mathcal{P}_1} e^{-\frac{\epsilon^2 nt}{48|S|}} \leq 2 \cdot O\left(2^{2d}\right) \cdot e^{-1.5\epsilon^2 t} \ll 1.$$

Thus A is an ϵ_0-approximation for $\mathcal{P}_1$.

Case $S \in \mathcal{P}_k \setminus \mathcal{P}_1$**:** For any set $S \in \mathcal{P}_k$, A is an ϵ'-approximation (Lemma 7.3), where

$$\epsilon' \leq \epsilon_0 + \sum_{i=1}^{k} 2\epsilon_i = \frac{\epsilon}{4} + \sum_{i=1}^{k} 2\frac{\sqrt{i}}{4 \cdot 2^i} \cdot \epsilon = \frac{\epsilon}{4} + \epsilon \cdot \sum_{i=1}^{k} \frac{\sqrt{i}}{2 \cdot 2^i} \leq \frac{\epsilon}{3}.$$

Case $S \in \mathcal{F} \setminus \mathcal{P}_k$**:** Let $F_S \in \mathcal{P}_k$ be such that $|\Delta\left(S, F_S\right)| \leq \frac{n}{2^{2k}} \leq \epsilon n$. Then all the $O\left(\frac{d}{\epsilon^2}\log\frac{1}{\epsilon}\right)$ sets in

$$\mathcal{R} = \{S - F_S, F_S - S \colon S \in \mathcal{F} \setminus \mathcal{P}_k\}$$

have size at most ϵn, and as shown before in equality (7.3), with non-zero probability, A is an $\frac{\epsilon}{3}$-approximation for $\mathcal{R}^{\dagger}$. As each set in $\mathcal{F} \setminus \mathcal{P}_k$ can be written as a set of $\mathcal{P}_k$

[†] Note that we only need to worry about over-sampling from each S of size at most ϵn.

with addition and substraction of two sets of $\mathcal{R}$, we conclude that A is an $\epsilon' + \frac{\epsilon}{3} + \frac{\epsilon}{3} \leq \epsilon$ approximation for all sets of $\mathcal{F} \setminus \mathcal{P}_k$.

This completes the proof.

Bibliography and discussion. The main theorem was proven in [1]. The proof given in the text is a simplification of the original proof by Mustafa and Ray.

[1] Y. Li, P. M. Long, and A. Srinivasan. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62(3):516–527, 2001.