CS6781 - Theoretical Foundations of Machine Learning

Lecture 8: Sample Complexity of Agnostic Learning

13 February 2020

Lecturer: Nika Haghtalab *Readings: Chapter 28.2, UML*
Scribes: Matt Peroni & Gregory Yauney

1 Overview

In the previous lecture, we discussed how one can relax the assumption of realizability in PAC learning and introduced the model of Agnostic PAC learning. In this lecture, we study the sample complexity of learning in the agnostic setting.

Definition 1.1 (Agnostic PAC learning)**.** An algorithm $\mathcal{A}$ agnostically PAC learns a class $\mathcal{C}$ if there exists a function $m_{\mathcal{C}}(\epsilon, \delta)$ such that the following is true: for any joint distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, and $\epsilon > 0$, any $\delta > 0$, and an i.i.d. sample set S of size $m \geq m_{\mathcal{C}}(\epsilon, \delta)$, $\mathcal{A}(S)$ returns a hypothesis $h_{\mathcal{A}}$ such that with probability $1 - \delta$:

$$\mathrm{err}_{\mathcal{D}}(h_{\mathcal{A}}) \leq \min_{c \in \mathcal{C}} \mathrm{err}_{\mathcal{D}}(c) + \epsilon$$

Definition 1.2 (Empirical risk minimization (ERM))**.** An algorithm $\mathcal{A}$ performs empirical risk minimization on hypothesis class $\mathcal{C}$ if, given a dataset S, $\mathcal{A}$ returns a hypothesis $h_{\mathcal{A}} \in \mathcal{C}$ that minimizes empirical error on the dataset:

$$h_{\mathcal{A}} \in \arg\min_{h \in \mathcal{C}} \mathrm{err}_S(h)$$

2 Sample complexity upper bound for finite $\mathcal{C}$

Theorem 2.1. *For any class $\mathcal{C}$, if algorithm $\mathcal{A}$ is an empirical risk minimizer, then $\mathcal{A}$ learns $\mathcal{C}$ with sample complexity:*

$$m_{\mathcal{C}}(\epsilon, \delta) = \frac{2}{\epsilon^2}\left(\ln\left(|\mathcal{C}|\right) + \ln\left(\frac{2}{\delta}\right)\right)$$

Definition 2.2 (Uniform Convergence)**.** Estimators $\mathrm{err}_S(\cdot)$ are said to have the uniform convergence property with respect to $\mathrm{err}_{\mathcal{D}}(\cdot)$ over the class of hypothesis $\mathcal{C}$, if for any ϵ and δ there is a function $m^{\mathrm{UC}}_{\mathcal{C}}(\epsilon, \delta) : (0,1) \times (0,1) \to \mathbb{N}$ such that for any $m \geq m^{\mathrm{UC}}_{\mathcal{C}}(\epsilon, \delta)$,

$$\Pr_{S \sim \mathcal{D}^m}\left[\exists h \in \mathcal{C}, |\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| \geq \epsilon\right] \leq \delta.$$

That is, the gap between true and empirical error of all hypothesis in $\mathcal{C}$ is bounded by ϵ with high probability.

We show that uniform convergence is sufficient for agnostic learning.

Proof of Theorem 2.1. The proof is almost exactly as in the realizable PAC setting. We care about the probability that $\mathrm{err}_{\mathcal{D}}(h)$ is large but $\mathrm{err}_S(h)$ is small. Fix a hypothesis $h \in \mathcal{C}$. If we can bound the probability of too large a gap between true and empirical errors for this fixed hypothesis, then we can use the union bound to bound the probability that any hypothesis in $\mathcal{C}$ has too large a gap.

That is, if we can show for a fixed h:

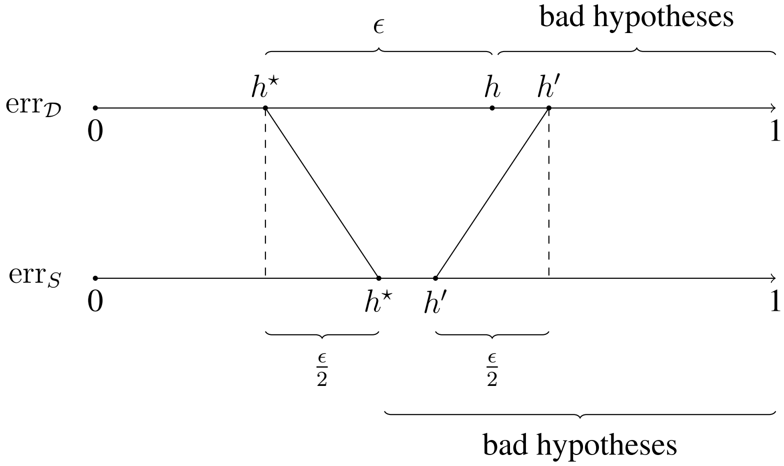
$$\Pr\left[|\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| \geq \frac{\epsilon}{2}\right] \leq 2\exp\left(\frac{-m\epsilon^2}{2}\right) \tag{1}$$

Then we can apply the union bound across all hypotheses in $\mathcal{C}$ to show:

$$\Pr\left[\exists h \in \mathcal{C} \text{ s.t. } |\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| \geq \frac{\epsilon}{2}\right] \leq 2\,|\mathcal{C}|\exp\left(\frac{-m\epsilon^2}{2}\right)$$

This is the same as saying $\mathcal{C}$ has uniform convergence property with parameters $\frac{\epsilon}{2}$ and δ. So what we are really showing is the following claim.

Claim 2.3. *Uniform convergence with parameters $\frac{\epsilon}{2}, \delta \implies$ ERM agnostically PAC learns $\mathcal{C}$ with parameters ϵ, δ.*



Let $h^\star \in \arg\min_{h\in\mathcal{C}} \mathrm{err}_{\mathcal{D}}(h)$ be an optimal classifier and let h be a hypothesis with mimimum $\mathrm{err}_S(h)$ returned by ERM. Now let h' be any classifier such that $\mathrm{err}_{\mathcal{D}}(h') > \mathrm{err}_{\mathcal{D}}(h^\star) + \epsilon$. Then, by uniform convergence $\mathrm{err}_S(h') > \mathrm{err}_{\mathcal{D}}(h') - \frac{\epsilon}{2}$. Moreover, $\mathrm{err}_S(h^\star) < \mathrm{err}_{\mathcal{D}}(h^\star) + \frac{\epsilon}{2}$. Therefore, $\mathrm{err}_S(h') > \mathrm{err}_S(h^\star)$, so h' cannot be the returned by the ERM. We can see this visually in the figure above.

So, all that is left to do is to prove Equation (1) above. We use the Hoeffding bound. In order to do so, we must define the error in terms of a sum of independent random variables.

Define a random variable for each example in S: $X_1, \ldots X_m$, where $X_i = \begin{cases} 1 & \text{if } h \text{ misclassifies } x_i \\ 0 & \text{otherwise} \end{cases}$

We can relate these independent random variables to the errors:

$$X = \frac{1}{m}\sum_{i=1}^{m} X_i = \frac{1}{m} \cdot \text{\# of samples } h \text{ misclassifies} = \mathrm{err}_S(h)$$
$$\mathbb{E}\left[X\right] = \mathrm{err}_{\mathcal{D}}(h)$$

Now apply the Hoeffding bound: $\Pr\left[\left|\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)\right| \geq \epsilon\right] \leq 2\exp\left(-2m\epsilon^2\right)$
Plugging in $\frac{\epsilon}{2}$:

$$\Pr\left[\left|\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)\right| \geq \frac{\epsilon}{2}\right] \leq 2\exp\left(\frac{-2m\epsilon^2}{4}\right) = 2\exp\left(\frac{-m\epsilon^2}{2}\right)$$

Bounding this probability by δ gives us the value of m in the claim. □

3 Sample complexity upper bound for infinite $\mathcal{C}$

The following theorem upper bounds the sample complexity of agnostic learning for infinite hypothesis classes in terms of the VC dimension.

Theorem 3.1. *ERM agnostically PAC earns $\mathcal{C}$ with sample complexity:*

$$m_{\mathcal{C}}(\epsilon, \delta) \in O\left(\frac{1}{\epsilon^2}\left(\mathrm{VCDim}\left(\mathcal{C}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right)$$

The proof of this theorem relies on a technique called “chaining”, which we will not cover in this class. Instead we will show the following weaker bound in this lecture:

$$m_{\mathcal{C}}(\epsilon, \delta) \in O\left(\frac{1}{\epsilon^2}\left(\ln\left(\Pi_{\mathcal{C}}\left(2m\right)\right) + \ln\left(\frac{1}{\delta}\right)\right)\right)$$

By Sauer’s lemma, this gives us:

$$m_{\mathcal{C}}(\epsilon, \delta) \in O\left(\frac{1}{\epsilon^2}\left(\mathrm{VCDim}\left(\mathcal{C}\right)\ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right)$$

Remark 3.2. This sample complexity in the agnostic setting has an additional factor of $\frac{1}{\epsilon}$ compared to the realizable setting.

Proof sketch. The proof is very similar to the sample complexity upper bound for realizable PAC. Therefore, we only highlight some of the major steps.

Take $S \sim \mathcal{D}^m$, and for the sake of analysis imagine an independent sample set $S' \sim \mathcal{D}^m$. Define the following “bad” events:

$$B(S) : \exists h \in \mathcal{C} \text{ such that } \left|\mathrm{err}_S(h) - \mathrm{err}_{\mathcal{D}}(h)\right| > \frac{\epsilon}{2}$$

$$B'(S,S') : \exists h \in \mathcal{C} \text{ such that } |\mathrm{err}_S(h) - \mathrm{err}_{S'}(h)| > \frac{\epsilon}{4}$$
$$B''(S,S',\vec{\sigma}) : \exists h \in \mathcal{C} \text{ such that } |\mathrm{err}_T(h) - \mathrm{err}_{T'}(h)| > \frac{\epsilon}{4}$$

In the final event, T and T' are defined by $\vec{\sigma} \in \{-1,+1\}^m$, which swaps elements of S and S':

$$\text{If } \sigma_i = +1 \rightarrow \begin{cases} \text{add } x_i \text{ to } T \\ \text{add } x_i' \text{ to } T' \end{cases} \qquad \text{If } \sigma_i = -1 \rightarrow \begin{cases} \text{add } x_i \text{ to } T' \\ \text{add } x_i' \text{ to } T \end{cases}$$

Just as in the proof of Theorem 1.1 of Lecture 5, it is sufficient to show that $\Pr_{S\sim\mathcal{D}^m}[B(S)] \leq \delta$.

Claim 3.3. *For large enough $m \in O(1/\epsilon^2)$,*

$$\Pr_{S\sim\mathcal{D}^m}[B'(S,S')|B(S)] \geq \frac{1}{2}$$

Proof. Take a hypothesis h such that $|\mathrm{err}_S(h) - \mathrm{err}_{\mathcal{D}}(h)| > \frac{\epsilon}{2}$. Then, by Jensen's inequality we have

$$\begin{aligned} \mathbb{E}_{S'\sim\mathcal{D}^m}[|\mathrm{err}_{S'}(h) - \mathrm{err}_S(h)|] &\geq |\mathbb{E}_{S'\sim\mathcal{D}^m}[\mathrm{err}_{S'}(h)] - \mathrm{err}_S(h)| \\ &\geq |\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| \\ &\geq \frac{\epsilon}{2} \end{aligned}$$

We want to show:

$$\Pr_{S'\sim\mathcal{D}^m}\left[|\mathrm{err}_S(h) - \mathrm{err}_{S'}(h)| - \mathbb{E}[|\mathrm{err}_S(h) - \mathrm{err}_{S'}(h)|] > \frac{\epsilon}{4}\right] \leq \frac{1}{2}$$

Note that $|\mathrm{err}_S(h) - \mathrm{err}_{S'}(h)|$ can be represented by the random variables: $\big|1\big(h(x_i') \neq y_i'\big) \neq 1\big(h(x_i) \neq y_i\big)\big|$, so we can use the Hoeffding bound for $m > \frac{1}{\epsilon^2}$ samples. □

This claim implies that $\Pr[B(S)] \leq 2\Pr[B'(S,S')]$, so it suffices to bound $\Pr[B'(S,S')]$.

Claim 3.4. *For i.i.d. sample sets $S \sim \mathcal{D}^m$ and $S' \sim \mathcal{D}^m$ and a vector $\vec{\sigma}$, where $\sigma_i = +1$ or -1 with probability 1/2 for all $i \in [m]$ independently, we have*

$$\Pr_{S,S'}[B'(S,S')] = \Pr_{S,S',\vec{\sigma}}[B''(S,S',\vec{\sigma})]$$

Proof. T, T' and S, S' are identically distributed. □

It now suffices to bound $\Pr[B''(S,S',\vec{\sigma})]$ in order to prove the theorem.

Claim 3.5. *For a fixed $S, S' \in \mathcal{X}^m$ and any h that is fixed (independently of $\vec{\sigma}$), we have*

$$\Pr_{\vec{\sigma}}\left[|\mathrm{err}_T(h) - \mathrm{err}_{T'}(h)| \geq \frac{\epsilon}{4} \mid S, S'\right] \leq 2\exp\left(\frac{-\epsilon^2 m}{8}\right)$$

Proof. We can rewrite our expression using indicator variables as

$$\mathrm{err}_T(h) - \mathrm{err}_{T'}(h) = \frac{1}{m}\sum_i \sigma_i(1(h(x_i) \neq y_i) - 1(h(x'_i) \neq y'_i))$$

Let us denote the term in the summation by z_i and let $z = \frac{1}{m}\sum_i z_i$. Recognizing that z_i is a random variable in $[-1, 1]$ with expected value 0, we can apply the Hoeffding bound to obtain the statement in the claim. That is,

$$\Pr\left[|z - \mathbb{E}[z]| \geq \frac{\epsilon}{4}\right] \leq 2\exp\left(\frac{-\epsilon^2 m}{8}\right)$$

as desired. □

With Claim 3.5, we can formulate the last step in our proof.

Claim 3.6.

$$\Pr\left[\exists h \in \mathcal{C} : |\mathrm{err}_T(h) - \mathrm{err}_{T'}(h)| \geq \frac{\epsilon}{4}|S, S'\right] \leq 2\Pi_{\mathcal{C}}(2m)\exp(\frac{-\epsilon^2 m}{8})$$

Proof. By the union bound, we have that

$$\begin{aligned}\Pr\left[\exists h \in \mathcal{C} : |\mathrm{err}_T(h) - \mathrm{err}_{T'}(h)| \geq \frac{\epsilon}{4}\middle| S, S'\right] &\leq \sum_{i=1}^{\Pi_{\mathcal{C}}(2m)} \Pr\left[|\mathrm{err}_T(h_i) - \mathrm{err}_{T'}(h_i)| \geq \frac{\epsilon}{4}\middle| S, S'\right] \\ &= 2\Pi_{\mathcal{C}}(2m)\exp\left(\frac{-\epsilon^2 m}{8}\right)\end{aligned}$$

Where we sum over $\Pi_{\mathcal{C}}(2m)$ hypotheses because that is the maximum possible number of unique hypotheses for $2m$ samples $S \cup S'$ in $\mathcal{C}$. The last line of the argument is from Claim 3.5. □

Bounding the probability in Claim 3.6 by $\delta/2$, we arrive at

$$m_{\mathcal{C}}(\epsilon, \delta) = \frac{8}{\epsilon^2}\left(\ln\left(\Pi_{\mathcal{C}}(2m)\right) + \ln\left(\frac{4}{\delta}\right)\right)$$

□

4 Agnostic PAC sample complexity lower bound

Theorem 4.1. *Any algorithm that agnostically PAC learns a hypothesis class $\mathcal{C}$ with VC-dimension d and parameters ϵ, δ has sample complexity:*

$$m \in \Omega\left(\frac{1}{\epsilon^2}\left(d + \ln\frac{1}{\delta}\right)\right)$$

We will prove the claim that when $\delta = \frac{1}{7}$, $m \geq \frac{d}{320\epsilon^2}$ samples are required.

4.1 High-level idea

Consider the following experiment. Imagine two coins.

$$\text{Coin 1:} \begin{cases} \frac{3}{4} & \text{H} \\ \frac{1}{4} & \text{T} \end{cases} \qquad \text{Coin 2:} \begin{cases} \frac{1}{4} & \text{H} \\ \frac{3}{4} & \text{T} \end{cases}$$

Choose one of these coins uniformly at random and flip that coin N times and see how many of these turns out heads and how many tails. What's the best strategy for guessing which coin was chosen? It is not hard to see that the best strategy is to guess coin 1 if more than half of the coin tosses are heads, and guess coin 2 otherwise. Indeed this strategy is best even for the following two coins.

$$\text{Coin 1:} \begin{cases} \frac{1+\epsilon}{2} & \text{H} \\ \frac{1-\epsilon}{2} & \text{T} \end{cases} \qquad \text{Coin 2:} \begin{cases} \frac{1-\epsilon}{2} & \text{H} \\ \frac{1+\epsilon}{2} & \text{T} \end{cases}$$

As ϵ approaches 0, the chance that our rule guesses the correct coin gets worse. That is because the probability that coin 1 is chosen but less than half of the coin tosses are heads (and coin 2 is chosen but more than half of the coin tosses are heads) increases for a fixed N as $\epsilon \to 0$. How many coin flips N are sufficient to guess correctly with fixed probability δ? Hoeffding bound dictates that $N = O\left(\frac{1}{\epsilon^2}\ln\left(\frac{1}{\delta}\right)\right)$ is sufficient: For coin 1,

$$\Pr\left[\left(\frac{1+\epsilon}{2}\right) - \frac{\epsilon}{2} > \frac{\#\text{ heads}}{N}\right] \leq \exp\left(\frac{-N\epsilon^2}{4}\right) \leq \delta.$$

As it turns out, Hoeffding bound is tight when the expected value $\mathbb{E}[X]$ is close to $\frac{1}{2}$. This is shown by the following inequality:

$$\Pr\left[\left(\frac{1+\epsilon}{2}\right) - \frac{\epsilon}{2} > \frac{\#\text{ heads}}{N}\right] > \frac{1}{2}\left(1 - \sqrt{1 - \exp\left(\frac{-N\epsilon^2}{1-\epsilon^2}\right)}\right) \qquad \text{(Slud's Inequality)}$$

So if $N < \frac{1-\epsilon^2}{\epsilon^2}$ then there is a constant probability of guessing incorrectly. We will use this to construct an example where the best learner has to see many samples to get a small error.

4.2 Lower bound

Take the set $Z = \{x_1, \ldots, x_d\}$ shattered by $\mathcal{C}$. The following class of distributions have the property that Bayes optimal classifier is almost as bad as any: For any $\vec{b} \in \{-1, 1\}^d$

$$\mathcal{D}_{\vec{b}} : \text{Has marginal distribution that is uniform on } \mathcal{X} \text{ and} \Pr[y_i|x_i] = \begin{cases} \frac{1+\rho}{2} & y_i = b_i \\ \frac{1-\rho}{2} & y_i = -b_i \end{cases}$$

The Bayes optimal classifier is $h^\star_{\vec{b}}(x_i) = b_i$ because the label of x_i is more likely to be b_i than $-b_i$.

For any $h \in \mathcal{C}$ and any $\vec{b}$:

$$\mathrm{err}_{\mathcal{D}_{\vec{b}}}(h) - \mathrm{err}_{\mathcal{D}_{\vec{b}}}(h^\star_{\vec{b}}) = \frac{1}{d} \cdot \frac{2\rho}{2} \cdot \left| \{i : h(x_i) \neq h^\star_{\vec{b}}(x_i)\} \right|$$

When $h(x_i) \neq y_i$ and $h^\star_{\vec{b}}(x_i) = y_i$, when h is incorrect but the Bayes optimal classifier is correct, there is an ρ difference in error. Samples are independent so seeing others gives no extra knowledge.

Claim 4.2 (With no proof). *Choose a $\mathcal{D}_{\vec{b}}$ by taking $b_i \sim$ Unif$(\{-1, +1\})$ independently. The algorithm $\mathcal{A}(\cdot)$ with best error in expectation over $\vec{b}$ and S, returns a classifier h such that $h(x_i) = +1$ if more than half of the x_i that appear in S are labeled $+1$ and -1 otherwise.*

Proof ideas. This can be proved by repeating the above coin experiments. Since b_i's are chosen independently (and Z is shattered by $\mathcal{C}$) what the optimal $\mathcal{A}(S)$ predicts for x_i should be independent of what it predicts for x_j. □

We want to show that

$$\mathbb{E}_{\vec{b}\sim\text{Unif}}\, \mathbb{E}_{S\sim\mathcal{D}^m_{\vec{b}}} \left[\mathrm{err}_{\mathcal{D}_{\vec{b}}}(h) - \mathrm{err}_{\mathcal{D}_{\vec{b}}}(h^\star_{\vec{b}})\right] \geq \frac{\rho}{4}$$

Claim 4.3. *For any algorithm $\mathcal{A}$ if $m \leq \frac{d(1-\rho^2)\ln(4/3)}{\rho^2}$,*

$$\mathbb{E}_{\vec{b}\sim\textit{Unif}}\, \mathbb{E}_{S\sim\mathcal{D}^m_{\vec{b}}} \left[\mathrm{err}_{\mathcal{D}_{\vec{b}}}(h_{\mathcal{A}(S)}) - \mathrm{err}_{\mathcal{D}_{\vec{b}}}(h^\star_{\vec{b}})\right] \geq \frac{\rho}{4}.$$

Proof sketch. Claim 4.2 shows that rather than taking $\vec{b}$ uniformly and considering the expected behavior of the best algorithm, we can fix $\vec{b}$ and assume that the algorithm is as described in Claim 4.2. Let $p_S(x_i)$ be the fraction of x_i points that appear in S and are labeled $+1$.

$$\begin{aligned}
&\mathbb{E}_{\vec{b}}\, \mathbb{E}_{S\sim\mathcal{D}^m_{\vec{b}}} \left[\mathrm{err}_{\mathcal{D}_{\vec{b}}}(h_{\mathcal{A}(S)}) - \mathrm{err}_{\mathcal{D}_{\vec{b}}}(h^\star_{\vec{b}})\right] \\
&\quad \geq \frac{\rho}{d} \cdot \Pr_{S\sim\mathcal{D}^m_{\vec{b}}} \left[\{i \,|\, b_i = +1 \wedge p_S(x_i) < \frac{1}{2}\} \cup \{i \,|\, b_i = -1 \wedge p_S(x_i) \geq \frac{1}{2}\}\right] \\
&\quad \geq \frac{\rho}{2}\left(1 - \sqrt{1 - \exp\left(\frac{-\,(m/d)\,\rho^2}{1-\rho^2}\right)}\right),
\end{aligned}$$

Where last inequality states (without proof) that best case scenario for the algorithm is if each instance x_i gets an equal $\frac{m}{d}$ of the samples and then uses the probability of the bad event in Section 4.1, the coin experiment. This is $\geq \frac{\rho}{4}$ if $m \leq \frac{d(1-\rho^2)\ln(4/3)}{\rho^2}$. □

Claim 4.4. *For any $\rho \leq \frac{1}{2}$ and for every algorithm $\mathcal{A}$, there is a distribution $\mathcal{D}$ such that if $m \leq \frac{d}{5\rho^2}$,*

$$\Pr_{S\sim\mathcal{D}^m}\left[\mathrm{err}_{\mathcal{D}}(h_{\mathcal{A}(S)}) - \min_{h\in\mathcal{C}} \mathrm{err}_{\mathcal{D}}(h) > \frac{\rho}{8}\right] > \frac{1}{7}$$

Proof. The proof is the same as in realizable PAC. It is immediate to go from expectation over all $\mathcal{D}_{\vec{b}}$ to existence of a bad distribution. Note that $m \leq \frac{d}{5\rho^2} \leq \frac{d(1-\rho^2)\ln(4/3)}{\rho^2}$. Let $\Delta = \mathrm{err}_{\mathcal{D}}(h_{\mathcal{A}(S)}) - \min_{h \in \mathcal{C}} \mathrm{err}_{\mathcal{D}}(h)$. Because $\Delta \in [0, \rho]$ and $\mathbb{E}[\Delta] \geq \frac{\rho}{4}$, we have $\Pr\left[\Delta > \frac{\rho}{8}\right] = p > \frac{1}{7}$:

$$\begin{aligned} \frac{\rho}{4} &\leq p\rho + (1-p)\frac{\rho}{8} \\ \implies p &\geq \frac{1}{7} \end{aligned}$$ □

Setting $\rho = 8\epsilon$ gives us the proof of Theorem 4.1.