# Appendix

# Proof of the VC Bound

In this Appendix, we present the formal proof of Theorem 2.5. It is a fairly elaborate proof, and you may skip it altogether and just take the theorem for granted, but you won't know what you are missing ☺ !

**Theorem A.1** (Vapnik, Chervonenkis, 1971).

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \le 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}.$$

This inequality is called the VC Inequality, and it implies the VC bound of Theorem 2.5. The inequality is valid for any target function (deterministic or probabilistic) and any input distribution. The probability is over data sets of size $N$. Each data set is generated *iid* (independent and identically distributed), with each data point generated independently according to the joint distribution $P(\mathbf{x}, y)$. The event $\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon$ is equivalent to the union over all $h \in \mathcal{H}$ of the events $|E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon$; this union contains the event that involves $g$ in Theorem 2.5. The use of the supremum (a technical version of the maximum) is necessary since $\mathcal{H}$ can have a continuum of hypotheses.

The main challenge to proving this theorem is that $E_{\text{out}}(h)$ is difficult to manipulate compared to $E_{\text{in}}(h)$, because $E_{\text{out}}(h)$ depends on the entire input space rather than just a finite set of points. The main insight needed to overcome this difficulty is the observation that we can get rid of $E_{\text{out}}(h)$ altogether because the deviations between $E_{\text{in}}$ and $E_{\text{out}}$ can be essentially captured by deviations between two in-sample errors: $E_{\text{in}}$ (the original in-sample error) and the in-sample error on a *second* independent data set (Lemma A.2). We have seen this idea many times before when we use a test or validation set to estimate $E_{\text{out}}$. This insight results in two main simplifications:

1. The supremum of the deviations over infinitely many $h \in \mathcal{H}$ can be reduced to considering only the dichotomies implementable by $\mathcal{H}$ on the

two independent data sets. That is where the growth function $m_{\mathcal{H}}(2N)$ enters the picture (Lemma A.3).

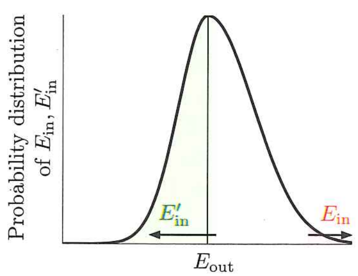
2. The deviation between two *independent* in-sample errors is 'easy' to analyze compared to the deviation between $E_{\text{in}}$ and $E_{\text{out}}$ (Lemma A.4).

The combination of Lemmas A.2, A.3 and A.4 proves Theorem A.1.

## A.1 Relating Generalization Error to In-Sample Deviations

Let's introduce a second data set $\mathcal{D}'$, which is independent of $\mathcal{D}$, but sampled according to the same distribution $P(\mathbf{x}, y)$. This second data set is called a *ghost* data set because it doesn't really exist; it is a just a tool used in the analysis. We hope to bound the term $\mathbb{P}[|E_{\text{in}} - E_{\text{out}}|$ is large] by another term $\mathbb{P}[|E_{\text{in}} - E'_{\text{in}}|$ is large], which is easier to analyze.

The intuition behind the formal proof is as follows. For any single hypothesis $h$, because $\mathcal{D}'$ is fresh, sampled independently from $P(\mathbf{x}, y)$, the Hoeffding Inequality guarantees that $E'_{\text{in}}(h) \approx E_{\text{out}}(h)$ with a high probability. That is, when $|E_{\text{in}}(h) - E_{\text{out}}(h)|$ is large, with a high probability $|E_{\text{in}}(h) - E'_{\text{in}}(h)|$ is also large. Therefore, $\mathbb{P}[|E_{\text{in}}(h) - E_{\text{out}}(h)|$ is large] can be approximately bounded by $\mathbb{P}[|E_{\text{in}}(h) - E'_{\text{in}}(h)|$ is large].




We are trying to bound the probability that $E_{\text{in}}$ is far from $E_{\text{out}}$. Let $E'_{\text{in}}(h)$ be the 'in-sample' error for hypothesis $h$ on $\mathcal{D}'$. Suppose that $E_{\text{in}}$ is far from $E_{\text{out}}$ with some probability (and similarly $E'_{\text{in}}$ is far from $E_{\text{out}}$, with that same probability, since $E_{\text{in}}$ and $E'_{\text{in}}$ are identically distributed). When $N$ is large, the probability is roughly Gaussian around $E_{\text{out}}$, as illustrated in the figure to the right. The red region represents the cases when $E_{\text{in}}$ is far from $E_{\text{out}}$. In those cases, $E'_{\text{in}}$ is far from $E_{\text{in}}$ about half the time, as illustrated by the green region. That is, $\mathbb{P}[|E_{\text{in}} - E_{\text{out}}|$ is large] can be approximately bounded by $2\,\mathbb{P}\,[|E_{\text{in}} - E'_{\text{in}}|$ is large].

This argument provides some intuition that the deviations between $E_{\text{in}}$ and $E_{\text{out}}$ can be captured by the deviations between $E_{\text{in}}$ and $E'_{\text{in}}$. The argument can be carefully extended to multiple hypotheses.

**Lemma A.2.**

$$\left(1 - 2e^{-\frac{1}{2}\epsilon^2 N}\right)\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \leq \mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \tfrac{\epsilon}{2}\right],$$

where the probability on the RHS is over $\mathcal{D}$ and $\mathcal{D}'$ jointly.





*Proof.* We can assume that $\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] > 0$, otherwise there is nothing to prove.

$$\begin{aligned}
&\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \tfrac{\epsilon}{2}\right] \\
\geq\ &\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \tfrac{\epsilon}{2} \textbf{ and } \sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \quad \text{(A.1)} \\
=\ &\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \times \\
&\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \tfrac{\epsilon}{2} \,\middle|\, \sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right].
\end{aligned}$$

Inequality (A.1) follows because $\mathbb{P}[\mathcal{B}_1] \geq \mathbb{P}[\mathcal{B}_1 \textbf{ and } \mathcal{B}_2]$ for any two events $\mathcal{B}_1, \mathcal{B}_2$. Now, let's consider the last term:

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \tfrac{\epsilon}{2} \,\middle|\, \sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right].$$

The event on which we are conditioning is a set of data sets with non-zero probability. Fix a data set $\mathcal{D}$ in this event. Let $h^*$ be any hypothesis for which $|E_{\text{in}}(h^*) - E_{\text{out}}(h^*)| > \epsilon$. One such hypothesis must exist given that $\mathcal{D}$ is in the event on which we are conditioning. The hypothesis $h^*$ does not depend on $\mathcal{D}'$, but it does depend on $\mathcal{D}$.

$$\begin{aligned}
&\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \tfrac{\epsilon}{2} \,\middle|\, \sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \\
\geq\ &\mathbb{P}\left[|E_{\text{in}}(h^*) - E'_{\text{in}}(h^*)| > \tfrac{\epsilon}{2} \,\middle|\, \sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \quad \text{(A.2)} \\
\geq\ &\mathbb{P}\left[|E'_{\text{in}}(h^*) - E_{\text{out}}(h^*)| \leq \tfrac{\epsilon}{2} \,\middle|\, \sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \quad \text{(A.3)} \\
\geq\ &1 - 2e^{-\frac{1}{2}\epsilon^2 N}. \quad \text{(A.4)}
\end{aligned}$$

1. Inequality (A.2) follows because the event "$|E_{\text{in}}(h^*) - E'_{\text{in}}(h^*)| > \frac{\epsilon}{2}$" implies "$\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2}$".
2. Inequality (A.3) follows because the events "$|E'_{\text{in}}(h^*) - E_{\text{out}}(h^*)| \leq \frac{\epsilon}{2}$" and "$|E_{\text{in}}(h^*) - E_{\text{out}}(h^*)| > \epsilon$" (which is given) imply "$|E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2}$".
3. Inequality (A.4) follows because $h^*$ is fixed with respect to $\mathcal{D}'$ and so we can apply the Hoeffding Inequality to $\mathbb{P}[|E'_{\text{in}}(h^*) - E_{\text{out}}(h^*)| \leq \frac{\epsilon}{2}]$.

Notice that the Hoeffding Inequality applies to $\mathbb{P}[|E'_{\text{in}}(h^*) - E_{\text{out}}(h^*)| \leq \frac{\epsilon}{2}]$ for any $h^*$, as long as $h^*$ is fixed with respect to $\mathcal{D}'$. Therefore, it also applies to any weighted average of $\mathbb{P}[|E'_{\text{in}}(h^*) - E_{\text{out}}(h^*)| \le \frac{\epsilon}{2}]$ based on $h^*$. Finally, since $h^*$ depends on a particular $\mathcal{D}$, we take the weighted average over all $\mathcal{D}$ in the event

$$\text{“} \sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\text{”}$$

on which we are conditioning, where the weight comes from the probability of the particular $\mathcal{D}$. Since the bound holds for every $\mathcal{D}$ in this event, it holds for the weighted average. ∎

Note that we can assume $e^{-\frac{1}{2}\epsilon^2 N} < \frac{1}{4}$, because otherwise the bound in Theorem A.1 is trivially true. In this case, $1 - 2e^{-\frac{1}{2}\epsilon^2 N} > \frac{1}{2}$, so the lemma implies

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \le 2\,\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2}\right].$$

## A.2 Bounding Worst Case Deviation Using the Growth Function

Now that we have related the generalization error to the deviations between in-sample errors, we can actually work with $\mathcal{H}$ restricted to two data sets of size $N$ each, rather than the infinite $\mathcal{H}$. Specifically, we want to bound

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2}\right],$$

where the probability is over the joint distribution of the data sets $\mathcal{D}$ and $\mathcal{D}'$. One equivalent way of sampling two data sets $\mathcal{D}$ and $\mathcal{D}'$ is to first sample a data set $S$ of size $2N$, then randomly partition $S$ into $\mathcal{D}$ and $\mathcal{D}'$. This amounts to randomly sampling, *without replacement*, $N$ examples from $S$ for $\mathcal{D}$, leaving the remaining for $\mathcal{D}'$. Given the joint data set $S$, let

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2} \,\middle|\, S\right]$$

be the probability of deviation between the two in-sample errors, where the probability is taken over the random partitions of $S$ into $\mathcal{D}$ and $\mathcal{D}'$. By the law of total probability (with $\sum$ denoting sum or integral as the case may be),

$$\begin{aligned}
&\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2}\right] \\
&= \sum_S \mathbb{P}[S] \times \mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2} \,\middle|\, S\right] \\
&\le \sup_S \ \mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2} \,\middle|\, S\right].
\end{aligned}$$





Let $\mathcal{H}(S)$ be the dichotomies that $\mathcal{H}$ can implement on the points in $S$. By definition of the growth function, $\mathcal{H}(S)$ cannot have more than $m_{\mathcal{H}}(2N)$ dichotomies. Suppose it has $M \le m_{\mathcal{H}}(2N)$ dichotomies, realized by $h_1, \ldots, h_M$. Thus,

$$\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| = \sup_{h\in\{h_1,\ldots,h_M\}} |E_{\text{in}}(h) - E'_{\text{in}}(h)|.$$

Then,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2} \,\middle|\, S\right] \\
&= \mathbb{P}\left[\sup_{h\in\{h_1,\ldots,h_M\}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2} \,\middle|\, S\right] \\
&\le \sum_{m=1}^{M} \mathbb{P}\left[|E_{\text{in}}(h_m) - E'_{\text{in}}(h_m)| > \frac{\epsilon}{2} \mid S\right] && \text{(A.5)} \\
&\le M \times \sup_{h\in\mathcal{H}} \mathbb{P}\left[|E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2} \mid S\right], && \text{(A.6)}
\end{aligned}$$

where we use the union bound in (A.5), and overestimate each term by the supremum over all possible hypotheses to get (A.6). After using $M \le m_{\mathcal{H}}(2N)$ and taking the sup operation over $S$, we have proved:

**Lemma A.3.**

$$\begin{aligned}
&\mathbb{P}\left[\sup_{h\in\mathcal{H}} |E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2}\right] \\
&\le m_{\mathcal{H}}(2N) \times \sup_S \sup_{h\in\mathcal{H}} \mathbb{P}\left[|E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2} \mid S\right],
\end{aligned}$$

where the probability on the LHS is over $\mathcal{D}$ and $\mathcal{D}'$ jointly, and the probability on the RHS is over random partitions of $S$ into two sets $\mathcal{D}$ and $\mathcal{D}'$.

The main achievement of Lemma A.3 is that we have pulled the supremum over $h \in \mathcal{H}$ outside the probability, at the expense of the extra factor of $m_{\mathcal{H}}(2N)$.

## A.3 Bounding the Deviation between In-Sample Errors

We now address the purely combinatorial problem of bounding

$$\sup_S \sup_{h\in\mathcal{H}} \mathbb{P}\left[|E_{\text{in}}(h) - E'_{\text{in}}(h)| > \frac{\epsilon}{2} \mid S\right],$$

which appears in Lemma A.3. We will prove the following lemma. Then, Theorem A.1 can be proved by combining Lemmas A.2, A.3 and A.4 taking $1 - 2e^{-\frac{1}{2}\epsilon^2 N} \ge \frac{1}{2}$ (the only case we need to consider).

**Lemma A.4.** For *any* $h$ and *any* $S$,

$$\mathbb{P}\left[|E_{\text{in}}(h) - E'_{\text{in}}(h)| > \tfrac{\epsilon}{2} \mid S\right] \leq 2e^{-\frac{1}{8}\epsilon^2 N},$$

where the probability is over random partitions of $S$ into two sets $\mathcal{D}$ and $\mathcal{D}'$.

*Proof.* To prove the result, we will use a result, which is also due to Hoeffding, for sampling *without replacement*:

**Lemma A.5** (Hoeffding, 1963). Let $\mathcal{A} = \{a_1, \ldots, a_{2N}\}$ be a set of values with $a_n \in [0, 1]$, and let $\mu = \frac{1}{2N}\sum_{n=1}^{2N} a_n$ be their mean. Let $\mathcal{D} = \{z_1, \ldots, z_N\}$ be a sample of size $N$, sampled from $\mathcal{A}$ uniformly *without* replacement. Then

$$\mathbb{P}\left[\left|\frac{1}{N}\sum_{n=1}^{N} z_n - \mu\right| > \epsilon\right] \leq 2e^{-2\epsilon^2 N}.$$

We apply Lemma A.5 as follows. For the $2N$ examples in $S$, let $a_n = 1$ if $h(\mathbf{x}_n) \neq y_n$ and $a_n = 0$ otherwise. The $\{a_n\}$ are the errors made by $h$ on $S$. Now randomly partition $S$ into $\mathcal{D}$ and $\mathcal{D}'$, i.e., sample $N$ examples from $S$ without replacement to get $\mathcal{D}$, leaving the remaining $N$ examples for $\mathcal{D}'$. This results in a sample of size $N$ of the $\{a_n\}$ for $\mathcal{D}$, sampled uniformly without replacement. Note that

$$E_{\text{in}}(h) = \frac{1}{N}\sum_{a_n \in \mathcal{D}} a_n, \quad \text{and} \quad E'_{\text{in}}(h) = \frac{1}{N}\sum_{a'_n \in \mathcal{D}'} a'_n.$$

Since we are sampling without replacement, $S = \mathcal{D} \cup \mathcal{D}'$ and $\mathcal{D} \cap \mathcal{D}' = \emptyset$, and so

$$\mu = \frac{1}{2N}\sum_{n=1}^{2N} a_n = \frac{E_{\text{in}}(h) + E'_{\text{in}}(h)}{2}.$$

It follows that $|E_{\text{in}} - \mu| > t \iff |E_{\text{in}} - E'_{\text{in}}| > 2t$. By Lemma A.5,

$$\mathbb{P}\left[|E_{\text{in}}(h) - E'_{\text{in}}(h)| > 2t\right] \leq 2e^{-2t^2 N}.$$

Substituting $t = \frac{\epsilon}{4}$ gives the result. ■