# Sample complexity for finite hypothesis classes

We discuss the connection between the PAC model and the CM (consistent hypothesis) model. We prove Valiant's sample complexity theorem, and briefly discusses its implication to *Occam's Razor*.

## 1 Valiant's Theorem

The PAC-learners for Boolean Conjunctions and Axis-Aligned Rectangles that we presented simply tried to find a hypothesis consistent with all the examples. It turns out that this is a general phenomenon: if the learner can produce a hypothesis consistent with "many" examples, then it is a PAC-learner. We first prove this simple fact for finite hypothesis classes. In particular, the following theorem was shown in the original paper by Les Valiant [2] which formulated PAC-learning and started the entire area of *computational learning theory* (COLT).

**Theorem 1.1** (Valiant's Theorem)**.** *If a learner can always produce a hypothesis consistent with*

$$m \geq \frac{1}{\epsilon} \log \left( \frac{|\mathcal{H}|}{\delta} \right)$$

*i.i.d. examples, then the learner is a PAC-learner.*

*Proof.* Call a hypothesis $h$ a "bad" hypothesis if $\text{err}_{\mathcal{D}}(h) > \epsilon$. The learner needs to output a good hypothesis with high confidence. Intuitively, a bad hypothesis is a hypothesis whose "disagreement region" with the target concept $c$ is large (larger than $\epsilon$).

Consider a fixed bad hypothesis $h$. Let $h \Delta c = \{x \in \Omega \mid h(x) \neq c(x)\}$ denote the "disagreement region" between $c$ and $h$. Then, saying that $h$ is bad is the same as saying that $\text{Prob}_{x \leftarrow \mathcal{D}}[x \in h\Delta c] > \epsilon$. Now, the crucial point to notice is this: if one of our examples belongs to the disagreement region, then $h$ will **not** be outputted because the output hypothesis is consistent with all examples. When we take $m$ samples, the probability that no sample hit the region $h\Delta c$ is at most $(1-\epsilon)^m$. Thus, the probability that $h$ is outputted is less than $(1-\epsilon)^m$. By the union bound, the probability that a bad hypothesis is outputted is at most $|\mathcal{H}|(1-\epsilon)^m \leq |\mathcal{H}|e^{-\epsilon m}$. This probability is at most $\delta$ if $m \geq \frac{1}{\epsilon}\log(|\mathcal{H}|/\delta)$. □

Intuitively, the theorem makes sense: if a hypothesis has large $\text{err}_{\mathcal{D}}(h)$, then its "error region" is large and so it is likely that some sample point will hit the error region and thus the hypothesis won't be outputted.

**Exercise 1.** Show that PAC-learning BOOLEAN CONJUNCTIONS only need $\frac{1}{\epsilon}(n \log 3 + \log(1/\delta))$ samples.

**Corollary 1.2.** *If learner can produce a hypothesis $h \in \mathcal{H}$ consistent with $m$ examples, then*

$$\textit{Prob}\left[ \textit{err}_{\mathcal{D}}(h) \leq \frac{1}{m} \log \left( \frac{|\mathcal{H}|}{\delta} \right) \right] \geq 1 - \delta$$

The corollary can be intuitively interpreted as follows. First, $\text{err}_{\mathcal{D}}(h)$ gets smaller when $m$ gets larger, because there's more data to learn from. Second, $\text{err}_{\mathcal{D}}(h)$ gets smaller when $|\mathcal{H}|$ gets smaller. The more we know about the concept, the smaller the hypothesis class becomes, thus the better the learning error.

## 2 Occam's Razor

The so-called Occam's Razor result was first observed in a short paper by Blumer et al. [1] in 1987. Although we do not know what Father William of Ockham (c. 1288 – c. 1348) really meant, in modern science Occam's Razor principle roughly refers to the philosophical position that "all else equal, the simpler hypothesis is the correct one."

Blumer et al. observed that Valiant's result can be used to give a mathematical justification to Occam's razor. More concretely, from Valiant's theorem we can prove that if the learner can always produce a "short" hypothesis consistent with the examples, then it is a PAC-learner. Here, we take the (information theoretic) view that "short = simple". Intuitively, the proof goes as follows: if the hypothesis is short, we can show that $\mathcal{H}$ is small, and thus the learner satisfies the theorem above. From now on, we will call a learner which always produce a consistent hypothesis an *Occam learner*.

**Theorem 2.1** (Occam's Razor, Roughly stated)**.** *If a learner always produce a hypothesis $h \in \mathcal{H}$ with $|h| = O((n|c|)^{\alpha} m^{\beta})$ for some fixed $\alpha$ (arbitrary) and $0 < \beta < 1$, then it is an efficient PAC-learner.*

## References


[1] A. BLUMER, A. EHRENFEUCHT, D. HAUSSLER, AND M. K. WARMUTH, *Occam's razor*, Inf. Process. Lett., 24 (1987), pp. 377–380. 2

[2] L. G. VALIANT, *A theory of the learnable*, Commun. ACM, 27 (1984), pp. 1134–1142. 1