Theory of Generalization Reading Group

Classical theory of generalization

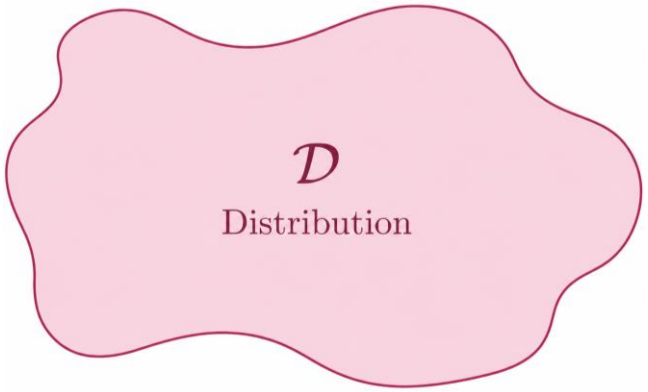
Based on: Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Outline

1. Why should training error imply test error?
2. PAC learning: what does it mean to learn?
3. No-Free-Lunch: why assumptions are necessary
4. VC dimension: measuring hypothesis complexity
5. Uniform convergence and the success of ERM
6. The Fundamental Theorem of Statistical Learning
7. From classical theory to modern deep learning

Training Error $\overset{?}{\Longrightarrow}$ Test Error?

The Learning Problem



We assume there exists an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$

and an unknown target function: $f : \mathcal{X} \to \{0, 1\}$.

We do **not** observe $\mathcal{D}$ directly.

But, we can sample from it: $S = \{(x_1, y_1), \ldots, (x_m, y_m)\} \sim \mathcal{D}^m$.

A learning algorithm *A* takes *S* and outputs $h = A(S)$.

Ideally, we would like to…

minimize the true risk $L_{\mathcal{D}}(h) = \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y]$ But we cannot compute it because $\mathcal{D}$ is unknown 😔

So instead, we minimize the empirical risk $L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}[h(x_i) \neq y_i].$

Now, the central question is: $L_S(h)$ small $\implies$ $L_{\mathcal{D}}(h)$ small ?

Under what conditions does minimizing the empirical risk lead to low population risk?

PAC Learning

What Does It Mean To Learn?

PAC (**P**robably **A**pproximately **C**orrect) Learning

We would like a learning algorithm that, given enough examples, outputs a hypothesis that performs well on unseen data.

But: How much data is enough, given an accuracy $1-\varepsilon$ and a confidence level $1-\delta$?

Definition: PAC Learnability

A hypothesis class H is **PAC learnable** if there exists a learning algorithm A and a sample complexity function

$$m_H(\varepsilon, \delta)$$

such that for every distribution $\mathcal{D}$, for every $\varepsilon, \delta > 0$, if

$$m \geq m_H(\varepsilon, \delta),$$

then with probability at least $1-\delta$, $\boxed{L_{\mathcal{D}}(A(S)) \leq \varepsilon.}$

A class of spam classification hypotheses is PAC-learnable, if for every level of accuracy $1-\varepsilon$ and confidence $1-\delta$, I can find a hypothesis **after seeing enough emails**

What If the True Function Is Not in Our Hypothesis Class?

Agnostic PAC Learning

The Realizability Assumption

In PAC Learning, we assume that the hypothesis class contains the target function f

Thus:

$$\mathrm{ERM}_H(S) \in \arg\min_{h \in H} L_S(h).$$

Or

$$L_S(\mathrm{ERM}_H(S)) = 0.$$

Hence assuming:

$$\boxed{L_{\mathcal{D}}(A(S)) \leq \varepsilon.}$$

If no Realizability is assumed, our goal becomes:

$$L_{\mathcal{D}}(A(S)) \leq \inf_{h \in H} L_{\mathcal{D}}(h) + \varepsilon$$

So a hypothesis class being Agnostic PAC Learnable means for every accuracy and confidence levels, there's a sample complexity:

$$m_H(\varepsilon, \delta)$$

Such that, the following is verified

$$L_{\mathcal{D}}(A(S)) \leq \inf_{h \in H} L_{\mathcal{D}}(h) + \varepsilon$$

Most real-world machine learning is agnostic.

Multi-class classification

Regression

Language modeling

The same framework extends by replacing the 0-1 loss with other loss functions.

$$\ell(y, \hat{y}) = (y - \hat{y})^2$$

$$\vdots$$

The No-Free Lunch Theorem

Without assumptions, learning is impossible…

The No-Free Lunch Theorem

Let $\mathcal{X}$ be the input domain with $|\mathcal{X}| > 2m.$

And S our training sample with m examples

Let $H_{\text{all}} = \{0,1\}^{\mathcal{X}}$ be the class of all binary classifiers (our hypothesis class).

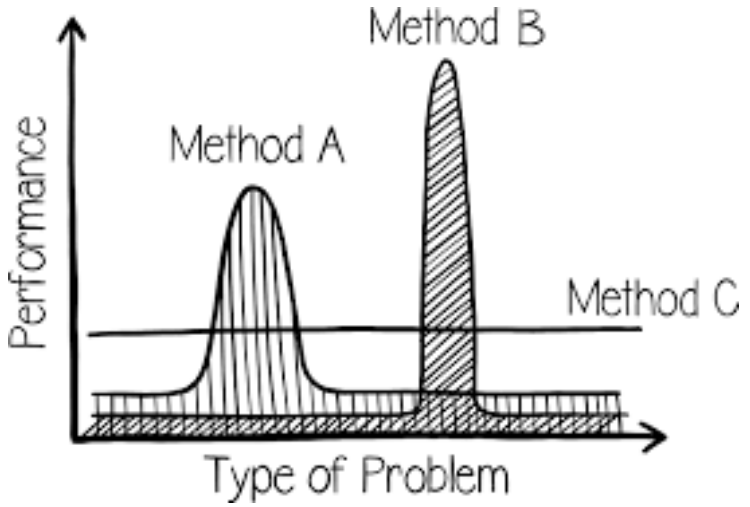
Then for every learning algorithm A, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:

1. $\mathcal{D}$ is realizable by H_{all}: $\exists h^* \in H_{\text{all}} \quad \text{s.t.} \quad L_{\mathcal{D}}(h^*) = 0.$

2. Yet:

$$\Pr_{S\sim\mathcal{D}^m}\left[L_{\mathcal{D}}(A(S)) \geq \frac{1}{8}\right] \geq \frac{1}{7}.$$

There exists no universal learner



For any unseen point $x \in C\backslash S$, the learner has no way to know which $h^*(x)$ is correct since both 0 and 1 are plausible by the hypothesis class

Escaping No-Free Lunch

What went wrong?

In No-Free-Lunch we used: $H_{\text{all}} = \{0,1\}^{\mathcal{X}}.$

Every labeling was possible.

$\Longrightarrow$

Consequence:

Observed labels gave no information about unseen labels.

Key Insight

Generalization requires **inductive bias**.

We need a hypothesis class that imposes structure.

Examples:

- Thresholds
- Linear classifiers
- Decision trees

So… Which hypothesis classes are learnable?

VC Dimension

Which Hypothesis Classes Are Learnable?

VC Dimension Measures Hypothesis Complexity

The No-Free-Lunch theorem failed because **every labeling was possible**.

VC dimension measures **how close a hypothesis class is to this situation**.

Restriction of H to a finite set: Consider $C = \{c_1, \ldots, c_m\}$.

The set of labelings induced by H is:

$$H_C = \{(h(c_1), \ldots, h(c_m)) : h \in H\}.$$

Definition (Shattering): We say that a hypothesis class shatters a finite set $C \subset \mathcal{X}$ if the restriction of H to C is the set of all functions from C to $\{0,1\}$. That is:

$$|\mathcal{H}_C| = 2^{|C|}$$

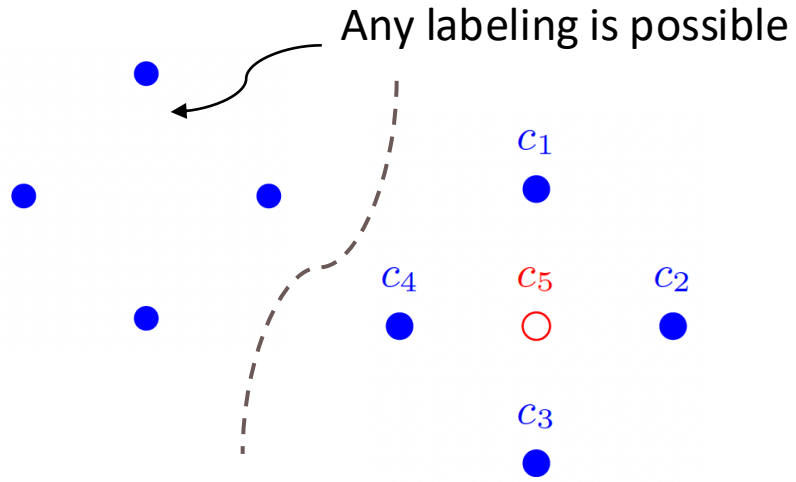
The VC-dimension of a hypothesis class H, denoted $VCdim(H)$, is the maximal size of a set $C \subset \mathcal{X}$ that can be shattered by H. If H can shatter sets of arbitrarily large size we say that H has infinite VC-dimension.

Example:

$$\mathcal{H} = \{h_{(a_1,a_2,b_1,b_2)} : a_1 \leq a_2 \text{ and } b_1 \leq b_2\}$$

Where:

$$h_{(a_1,a_2,b_1,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq a_2 \text{ and } b_1 \leq x_2 \leq b_2 \\ 0 & \text{otherwise} \end{cases}$$



$$\text{VCdim}(H) = 4.$$

Why should finite VC dimension imply learnability?

Uniform Convergence and ERM

Uniform Convergence

Finite VC dimension prevents the No-Free-Lunch phenomenon. But we still need to explain why it makes ERM work.

Finite VC dimension limits the effective size of the hypothesis class

For any fixed hypothesis $h \in H$, by Hoeffding's inequality:

$$\Pr(|L_D(h) - L_S(h)| > \varepsilon) \le e^{-cm\varepsilon^2}.$$

ERM chooses: $h_{ERM} = \arg\min_{h \in H} L_S(h)$, so we need $\forall h \in H$.

If H were finite: $\Pr(\exists h \in H : |L_D(h) - L_S(h)| > \varepsilon) \le |H|\, e^{-cm\varepsilon^2}$.

But what if H is infinite? A key observation is that on a finite sample, many hypotheses behave identically.

What matters is not | H |, but the number of distinct labelings | H_C |.

Spauer-Shelah-Perles Lemma

If: $VC(H) = d$, then $|H_C| \le \sum_{i=0}^{d} \binom{m}{i} \le \left(\frac{em}{d}\right)^d$.

$\Longrightarrow$ $\Pr(\exists h \in H : |L_D(h) - L_S(h)| > \varepsilon) \le \left(\frac{em}{d}\right)^d e^{-cm\varepsilon^2}$.

Therefore $\left(\frac{em}{d}\right)^d e^{-cm\varepsilon^2} \to 0$.

$\sup_{h \in H} |L_D(h) - L_S(h)| \to 0$.

$$VCdim < \infty \Rightarrow Uniform\ Convergence$$

The Fundamental Theorem of Statistical Learning

The Fundamental Theorem of Statistical Learning

All the notions we've introduced turn out to be equivalent

$$\boldsymbol{VCdim < \infty \Leftrightarrow Uniform\ Convergence \Leftrightarrow PAC\ Learnability \Leftrightarrow Agnostic\ PAC\ Learnability \Leftrightarrow ERM\ Succeeds}$$

Generalization comes from controlling the complexity of the hypothesis class

$$L_{\mathcal{D}}(A(S)) \leq \underbrace{\inf_{h \in H} L_{\mathcal{D}}(h)}_{\text{bias}} + \underbrace{\varepsilon}_{\text{complexity}}$$

Bias-complexity tradeoff

Format Statement (page 72)

THEOREM 6.7 (The Fundamental Theorem of Statistical Learning) *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ and let the loss function be the $0-1$ loss. Then, the following are equivalent:*

1. *$\mathcal{H}$ has the uniform convergence property.*
2. *Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$.*
3. *$\mathcal{H}$ is agnostic PAC learnable.*
4. *$\mathcal{H}$ is PAC learnable.*
5. *Any ERM rule is a successful PAC learner for $\mathcal{H}$.*
6. *$\mathcal{H}$ has a finite VC-dimension.*

From Classical Theory to Modern Deep Learning

Notes on Classical Theory and DL

Intuition

A smaller hypothesis class generalizes better because it cannot fit arbitrary labelings.

The Puzzle

Modern neural networks seem to violate this intuition.

- Millions (or billions) of parameters
- Extremely large VC dimension
- Can perfectly fit random labels
- Often interpolate the training data

Yet:

- Test error remains low
- Scaling often improves generalization

Goal of this Reading Group: **Understanding why overparameterized models generalize despite seemingly violating the classical picture.**