On the Mathematical Foundations of Learning: A Review

Aly El Gamal
ECE Department and Coordinated Science Laboratory
University of Illinois at Urbana-Champaign

Abstract

Based on [1], we present a review of key mathematical tools that aid in the solution and analysis of the problem of learning from examples. In the sequel, we present a linear vector space formulation, using which many of the results in this area (see e.g. [3]) apply, and in some cases, reflect a rather general concept. We also provide an upper bound analysis on the *generalization* error, during which, we decompose it into separate contributions of the sampling process and the choice of the hypothesis space. We finally study the problem of choosing the hypothesis space and, in particular, determining its optimal dimension.

I. Introduction

The problem of learning from examples is a commonly known problem that lies at the very core of the scientific foundation, and has both clear mathematical and practical interests. A famous example is that of acquiring the grammatical rules of a language by children, only through exemplary statements. The child can, from finitely many examples, with a probability that is rather high, deduce - with sufficient accuracy - the rules of her native language, that enables her to generalize the finite set of sampled sentences to an infinite one. In his book [2], Partha Niyogi states that "This povery of stimulus in the child language acquisition process motivated Chomsky to suggest that children operate with hypotheses about language which are constrained in some fashion.".

Figure 1 depicts the various factors that affects the learning problem. namely,

- **Concept Class**: The set of all possible solutions. For the language acquisition problem, this is the set of all possible languages, using which, humans may communicate.
- **Mechanism of Gathering Examples**: Samples may be drawn from a probability space with a defined measure, as will be considered later in the sequel. This can also used to model the sampling process for the language acquisition problem.
- **Noise**: Samples may not be *faithful* to truth. Either through noise in the sampling process, or the realization of low probability events.
- **Distance Metric**: A metric has to be defined to measure how *good* any given solution is.
- **Hypothesis Class**: The search space. This element is what Chomsky suggests to exist a priori when children learn the native language.
- **Learning Algorithm**: The learning algorithm, restricted to the hypothesis class, and based on the gathered examples and the error criterion defined by the distance metric, selects the best solution.

In a formal setting, various questions may arise in this problem. We list a few, which will be the subject of study of this article.

- Fixing all the parameters above but the last one, it is desired to find a learning algorithm that finds the best estimate. We tackle this question in Section III
- Fixing all the parameters of the problem, it is desired to obtain guarantees on the distance between the estimate and the correct solution. sections IV and V provide probabilistic upper bounds on the error for the problem defined in Section II
- Fixing all the parameters except the number of sample points, lower bounds on the number of samples that suffice to achieve specified low values for the error criterion are of interest. This easily follows from the analysis of the error bounds.
- Fixing all the parameters except the hypothesis class. It is interesting to study the various aspects governing the selection of such a class. This problem is highlighted in Section VI

Finally, we suggest future directions of research in Section VII, and conclude the article in Section VIII. We now proceed by providing a formal setting for the problem in the next Section.

II. Problem Definition

Here, we provide a rather general framework under which all the concepts highlighted in the sequel apply. Given two sets X, Y, the *goal* is to learn a function $f_{opt} : X \to Y$ from a finite set of data points $\mathbf{z} = \{z_1, z_2, \ldots, z_m\} : z_i = (x_i, y_i)$. Two properties of the sample points form an essential part of this problem. namely,

- The points $\{x_i, i \in [m]\}$ may be *sparse* with respect to the size of X.
- The images $\{y_i, i \in [m]\}$ may not *faithfully* describe the function f_{opt}.

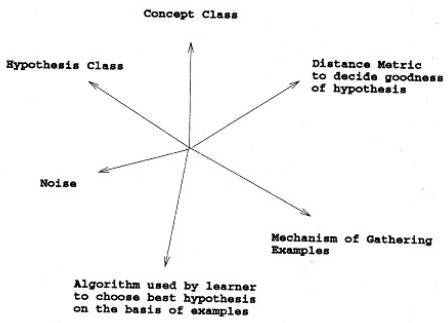


Fig. 1: Elements affecting the informational complexity of learning from examples

In order to well pose the problem, a norm has to be defined on the space of all functions from X to Y. Let $\hat{f}$ be the estimated function, then the error is given by.

$$\mathcal{E}(\hat{f}) = \|f_{opt} - \hat{f}\|^2 \tag{1}$$

Because of the items listed above, i.e. the samples are sparse and erroneous, as will be pointed out in more detail later in the sequel, a larger search space does not necessarily imply a more accurate solution. Hence, we restrict our attention to a subspace of the space of all functions from X to Y, which we call the hypothesis space $\mathcal{H}$. Consequently, we define our estimate $\hat{f}$ given the vector of samples $\mathbf{z}$ and the hypothesis space $\mathcal{H}$, as $\hat{f}_{\mathbf{z},\mathcal{H}}$, and is given by,

$$\hat{f}_{\mathbf{z},\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}_{\mathbf{z}}(f) \tag{2}$$

where the empirical error $\mathcal{E}_{\mathbf{z}}(f)$ is defined as,

$$\mathcal{E}_{\mathbf{z}}(f) = \sum_{i=1}^{m} \left(f(x_i) - y_i\right)^2 \tag{3}$$

We also define $\mathcal{E}_{\mathbf{z}}(\mathcal{H})$ as the best empirical error obtainable over $\mathcal{H}$. i.e.,

$$\mathcal{E}_{\mathbf{z}}(\mathcal{H}) = \mathcal{E}_{\mathbf{z}}\left(\hat{f}_{\mathbf{z},\mathcal{H}}\right) \tag{4}$$

Similarly, we define the best estimate $f_{\mathcal{H}}$, and the minimum error over $\mathcal{H}$, with respect to the error criterion imposed by the defined norm, as,

$$f_{\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}(f) \tag{5}$$

$$\mathcal{E}(\mathcal{H}) = \mathcal{E}\left(f_{\mathcal{H}}\right) \tag{6}$$

Finally, we define $\mathcal{E}_{\mathcal{H}}(f) = \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}})$. We call $\mathcal{E}(\mathcal{H})$ the *approximation* error, as it captures the part of the error that follows from approximating the space of functions by the hypothesis space $\mathcal{H}$. We note the following.

- The empirical estimate $\hat{f}_{\mathbf{z},\mathcal{H}}$ does not depend on the norm $\|.\|$ imposed on the space of functions, and hence, may **not** capture the minimum error $\mathcal{E}(f_{\mathcal{H}})$.
- The approximation error $\mathcal{E}(\mathcal{H})$ depends only on the imposed norm and the hypothesis space, and is independent from the sample points given by $\mathbf{z}$.

III. Linear Vector Space Formulation

In this Section, we derive a formulation of the solution to the above described problem, that makes use of the already available results on solutions to the following least square problem.

$$\min_{\mathbf{w}} \|A\mathbf{w} - \mathbf{y}\|^2 \tag{7}$$

where A is a matrix representation for a linear operator.

We restrict our attention here to the case where the hypothesis space $\mathcal{H}$ is a finite dimensional linear vector space. Let $\phi_1, \ldots, \phi_N$ be a basis for $\mathcal{H}$. Then, it follows by [[3], Theorem 4.1], that $\forall f \in \mathcal{H}, \exists \mathbf{w} \in \mathbf{R}^N$, such that,

$$f = \sum_{i=1}^{N} w_i \phi_i \tag{8}$$

For a set of samples $\mathbf{z}$ of size m, we know from the above definition of $f_{\mathbf{z},\mathcal{H}}$ that it minimizes over all functions $f \in \mathcal{H}$ the empirical error $\mathcal{E}_{\mathbf{z}}(f)$, and the minimum is $\mathcal{E}_{\mathbf{z}}(\mathcal{H})$, which is given by,

$$\begin{aligned}
\mathcal{E}_{\mathbf{z}}(\mathcal{H}) &= \min_{f\in\mathcal{H}} \sum_{j=1}^{m} (f(x_j) - y_j)^2 && (9)\\
&= \min_{\mathbf{w}\in\mathbf{R}^{\mathbf{N}}} \sum_{j=1}^{m} \left(\sum_{i=1}^{N} (w_i \phi_i(x_j)) - y_j \right)^2 && (10)\\
&= \min_{\mathbf{w}\in\mathbf{R}^{\mathbf{N}}} \sum_{j=1}^{m} \left(\sum_{i=1}^{N} (w_i a_{ij}) - y_j \right)^2 && (11)\\
&= \min_{\mathbf{w}\in\mathbf{R}^{\mathbf{N}}} \sum_{j=1}^{m} \left((A\mathbf{w})_j - y_j \right)^2 && (12)\\
&= \min_{\mathbf{w}\in\mathbf{R}^{\mathbf{N}}} \|A\mathbf{w} - \mathbf{y}\|^2 && (13)
\end{aligned}$$

where $a_{ij} = \phi_i(x_j)$. Hence, $A \in \mathbf{R}^{m\times N}$ is the matrix representation of the linear operator $\mathcal{A} : \mathbf{R}^N \to \mathbf{R}^m$ mapping functions in $\mathcal{H}$ to their evaluations at the X coordinates of the sample points in $\mathbf{z}$. As illustrated in [3], this problem has a unique solution if and only if $\mathcal{A}$ is injective, i.e. functions can be distinguished from their evaluations at the given sample points. In this case, A is full column rank and the solution is given by,

$$\mathbf{w} = A^{+}\mathbf{y} \tag{14}$$

where $A^{+} = \left(A^H A\right)^{-1} A^H$ is the Moore-Penrose pseudo inverse of A.

It is worth noting that in cases where more than one function in $\mathcal{H}$ can have the same evaluation at the considered data points, it is of interest to derive conditions under which the minimum norm solution (MNLS) gives a *good* answer with respect to the error criterion.

IV. Decomposition of Error

In this Section, we consider a more formal setting, where X is compact domain, $Y = \mathbf{R}$. We define the probability space Z by the product space $X \times Y$ with a Borel probability measure ρ. The only information that can be available about ρ is its marginal density ρ_X. Let $\mathcal{F}$ be the space of real valued random variables with probability measure ρ. Each function $f : X \to Y$ induces a random variable $f_Z \in \mathcal{F}$ in the Hilbert subspace of F defined on the probability space Z, where,

$$f_Z(x,y) = \begin{cases} y & \text{iff} \quad f(x) = y \\ 0 & \text{otherwise} \end{cases}$$

The norm of f is defined as $\|f_Z\|_{\mathcal{F}}$, which is given by,

$$\|f\| = \sqrt{E[f_Z^2]} \tag{15}$$

where for any random variable $\tilde{R} \in \mathcal{F}$, $E\left[\tilde{R}\right]$ denotes its expectation. i.e.,

$$E\left[f(x)\right] = \int_Z f(x) \quad d_\rho \tag{16}$$

It is clear that this defines a valid norm with respect to an equivalence relation that equates all random variables that are equal almost everywhere. We also assume that the sample points given by the vector $\mathbf{z}$ are drawn independently according to the probability measure ρ.

With a slight abuse of notation, we use the same symbols of X and Y, to denote the random variables that inherit the values of the corresponding sets with the probability measure ρ. It should be clear from context whether we mean the random variables or the spaces. Let $\mathcal{F}_X \subset \mathcal{F}$ be the subspace of random variables who have the same distribution as X. i.e., the subspace of all functions $f : X \to Y$. As in [3], we use $P_S s$ to denote the orthogonal projection of the element $s \in \mathcal{F}$ onto the subspace $S \subset \mathcal{F}$.

Theorem 1:

$$P_{\mathcal{F}_X} Y = E[Y|X] \tag{17}$$

Proof: It is clear that $\forall f \in \mathcal{F}_X, E[f|X] = f$. Also, the range space $\mathcal{R}(E[Y|X]) = \mathcal{F}_X$. Hence, it suffices to show that $(E[Y|X] - Y) \perp \mathcal{F}_X$. Indeed, $\forall f \in \mathcal{F}_X$,

$$\begin{aligned}
E\left[(E[Y|X] - Y)\, f\right] &= E\left[E[Y|X] f\right] - E[Y f] && (18)\\
&= E\left[E[Y f|X]\right] - E[Y f] && (19)\\
&= 0 && (20)
\end{aligned}$$

where the last equality is a consequence of the law of iterated expectation. The statement follows by the orthogonality principle. ■

As it will be commonly use, from now on, we denote $E[Y|X]$ as f_ρ. The next Theorem characterizes the error of a function $f : X \to Y$ by its distance in $\mathcal{F}$ from f_ρ.

Theorem 2:

$$\forall f \in \mathcal{F}_X, \mathcal{E}(f) = \|f - f_\rho\|^2 + \sigma_\rho^2 \tag{21}$$

Proof: We know from Theorem 1 that $f_\rho = P_{\mathcal{F}_X} Y$. Hence, $(Y - f_\rho) \perp \mathcal{F}_X$. Also, both $f, f_\rho \in \mathcal{F}_X$, so is $f - f\rho$. It follows by Pythagoras Theorem that,

$$\begin{aligned} \mathcal{E}(f) &= \|f - Y\|^2 & (22)\\ &= \|f - f_\rho\|^2 + \|f_\rho - Y\|^2 & (23) \end{aligned}$$

■

We note that the part of the error due to $\sigma_\rho^2 = \|f_\rho - Y\|^2$ does not depend on the choice of the estimate, hence, is inevitable. So, the goal of this problem is to find the function f in the hypothesis space space $\mathcal{H}$ that is closest to f_ρ. We next show, under specific conditions of $\mathcal{H}$, a decomposition of this distance, into a part that depends on the sample points, and another that depends on the choice of the hypothesis space.

Theorem 3: If $\mathcal{H}$ is a Banach space, then,

$$\forall f \in \mathcal{H}, \|f - f_\rho\|^2 = \|f_\mathcal{H} - f_\rho\|^2 + \|f - f_\mathcal{H}\|^2 \tag{24}$$

Proof: We know that if $\mathcal{H}$ is complete, then $f_\mathcal{H} = P_\mathcal{H} f_\rho$. Hence, $(f_\rho - f_\mathcal{H}) \perp (f - f_\mathcal{H}), \forall f \in \mathcal{H}$. Hence,

$$\|f - f_\rho\|^2 = \|f_\mathcal{H} - f_\rho\|^2 + \|f - f_\mathcal{H}\|^2 \tag{25}$$

■

In particular, the above result shows that $\mathcal{E}_\mathcal{H}(f) = \|f - f_\mathcal{H}\|^2, \forall f \in \mathcal{H}$, when $\mathcal{H}$ is a Banach space. We next relax the assumption on H and derive a more general relation between the relative error on $\mathcal{H}$ of a function and its distance to the best estimate $f_\mathcal{H}$. We use $\mathcal{L}_\rho^2(X)$ to denote the Hilbert space of all square integrable functions on X with respect to ρ.

Theorem 4 ([1], Chapter 1, Lemma 5): If $\mathcal{H}$ is convex and $f_\mathcal{H}$ exists, then $f_\mathcal{H}$ is unique as an element in $\mathcal{L}_\rho^2(X)$ and, $\forall f \in \mathcal{H}$,

$$\|f_\mathcal{H} - f\|^2 \leq \mathcal{E}_\mathcal{H}(f) \tag{26}$$

Proof: Assume that $f_\mathcal{H}$ exists. Let s be the line segment $\overline{f_\mathcal{H} f}$, then $s \subset \mathcal{H}$

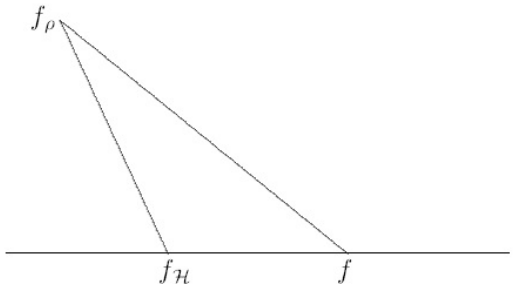


Fig. 2: $\|f_\rho - f_\mathcal{H}\|^2 \leq \|f_\rho - g\|^2, \forall g \in \overline{f_\mathcal{H} f} \Rightarrow \angle f_\rho f_\mathcal{H} f$ is obtuse

Now, $\forall g \in s$,

$$\|f_\rho - f_\mathcal{H}\|^2 \leq \|f_\rho - g\|^2 \tag{27}$$

Hence, $\angle f_\rho f_\mathcal{H} f$ is obtuse. It then follows that,

$$\begin{aligned} \|f_\mathcal{H} - f\|^2 &\leq \|f_\rho - f\|^2 - \|f_\rho - f_\mathcal{H}\|^2 & (28)\\ &= \mathcal{E}_\mathcal{H}(f) & (29) \end{aligned}$$

For proof of uniqueness of $f_\mathcal{H}$. Let f_1 be another minimizer of the distance to f_ρ in $\mathcal{L}_\rho^2(X)$. then both $\angle f_\rho f_\mathcal{H} f_1$ and $\angle f_\rho f_1 f_\mathcal{H}$ are obtuse. This implies that $f_1 = f_\mathcal{H}$. ■

V. Upper Bound on Error

In this Section, we provide conditions, under which upper bounds can be derived on the error of the best empirical estimate $\hat{f}_{\mathbf{z},\mathcal{H}}$. First, we recall that,

$$\mathcal{E}(\hat{f}_{\mathbf{z},\mathcal{H}}) = \mathcal{E}_{\mathcal{H}}(\hat{f}_{\mathbf{z},\mathcal{H}}) + \mathcal{E}(f_{\mathcal{H}}) \tag{30}$$

which follows from the definition of $\mathcal{E}_{\mathcal{H}}(f)$ for function $f \in \mathcal{H}$. The advantage of considering this decomposition, is the separation of the contributions of the sampling process and the selection of the hypothesis space, to the overall error. As will be pointed out later, this also highlights the trade-off between both parts as we increase the size of $\mathcal{H}$. Due to obvious reasons, we call $\mathcal{E}_{\mathcal{H}}(\hat{f}_{\mathbf{z},\mathcal{H}})$ the sample error, and $\mathcal{E}(\mathcal{H}) = \mathcal{E}(f_{\mathcal{H}})$ the approximation error.

A. Bounding the Sample Error

Here, we provide upper bounds on the sample error, noting that we assume an algorithm that successfully finds the estimate $\hat{f}_{\mathbf{z},\mathcal{H}}$, as we do not consider this problem here. With a slight loss in generality, we apply regularity properties on f_ρ and $\mathcal{H}$, that is needed for the proofs below. More specifically, we restrict our attention to the space $\mathcal{C}_X$ of continuous functions $f : X \to Y$. Also, we assume that $\mathcal{H} \subset \mathcal{C}_X$ is compact, and note that the infinity norm $\|.\|_\infty$ is well defined on both $\mathcal{C}_X$ and $\mathcal{H}$. Recall, that a space is compact, if and only if for every covering with open sets, there exists a finite sub-covering. i.e., for any collection of open sets $\{U_\alpha\}_{\alpha \in A}$ such that,

$$\mathcal{H} = \cup_{\alpha \in A} U_\alpha \tag{31}$$

there exists a finite subset $J \subset A$, such that,

$$\mathcal{H} = \cup_{\alpha \in J} U_\alpha \tag{32}$$

Define the *defect* $L_{\mathbf{z}}(f)$ as the difference between the real and empirical errors of a function $f \in \mathcal{F}_X$,

$$L_{\mathbf{z}}(f) = \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f) \tag{33}$$

Now, as the error $\mathcal{E}(f)$ is unknown, $|L_{\mathbf{z}}(f)|$ indicates how accurate the available empirical error approximates the real one. The following Theorem bounds the sample error, when a *uniform* estimate on the defect is present. i.e., for a good enough sampling process.

Theorem 5: For $\epsilon > 0$, $0 < \delta < 1$, if the following is true.

$$\text{Prob}_{\mathbf{z} \in Z^m} \left\{ \sup_{f \in \mathcal{H}} |L_{\mathbf{z}}(f)| \leq \epsilon \right\} \geq 1 - \delta \tag{34}$$

then,

$$\text{Prob}_{\mathbf{z} \in Z^m} \left\{ \mathcal{E}_{\mathcal{H}}(\hat{f}_{\mathbf{z},\mathcal{H}}) \leq 2\epsilon \right\} \geq 1 - \delta \tag{35}$$

Proof: It is sufficient to show that,

$$\mathcal{E}_{\mathcal{H}}(\hat{f}_{\mathbf{z},\mathcal{H}}) \leq 2\epsilon \Rightarrow \sup_{f \in \mathcal{H}} |L_{\mathbf{z}}(f)| \leq \epsilon \tag{36}$$

To this end, we note that,

$$\begin{aligned} \mathcal{E}(\hat{f}_{\mathbf{z},\mathcal{H}}) &\overset{(a)}{\leq} \mathcal{E}_{\mathbf{z}}(\hat{f}_{\mathbf{z},\mathcal{H}}) + \epsilon & (37)\\ &\overset{(b)}{\leq} \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) + \epsilon & (38)\\ &\overset{(c)}{\leq} \mathcal{E}(f_{\mathcal{H}}) + 2\epsilon & (39) \end{aligned}$$

where (a) and (c) follow from the uniform estimate on $|L_{\mathbf{z}}(f)|$, for all $f \in \mathcal{H}$, (b) follows from the fact that $f_{\mathbf{z},\mathcal{H}}$ minimizes the empirical estimate over all functions in $f \in \mathcal{H}$. ■

In case the statement of the above Theorem holds, the empirical estimate $\hat{f}_{\mathbf{z},\mathcal{H}}$ approximates the best estimate $f_{\mathcal{H}}$ with high probability. Hence, this solution is named *probably approximately correct*, or PAC.

We now provide the condition in the above Theorem. We first start by deriving a *pointwise* bound on the defect. i.e., bounding $L_{\mathbf{z}}(f)$ at each f. Recall that the role of the probability measure ρ in the considered problem is two-fold.

- The error function $\mathcal{E}(f)$ depends on ρ.
- The sample points in $\mathbf{z}$ are drawn independently according to ρ.

Hence, though the empirical error does not explicitly depend on ρ, but it does, through the samples. We know from the law of large numbers that $\forall f \in \mathcal{F}_X$,

$$\mathcal{E}_{\mathbf{z}}(f) \to \mathcal{E}(f), \text{ as } m \to \infty \tag{40}$$

Hence, $L_{\mathbf{z}}(f) \to 0$. The following inequality, introduced by Bernstein characterizes the convergence rate, hence, giving an upper bound on $|L_{\mathbf{z}}(f)|$ for a fixed number of samples.

Theorem 6: for $\zeta \in \mathcal{F}$, if $\exists M > 0$ such that, $|\zeta(x,y) - E[\zeta]| \leq M$ almost everywhere on $Z = X \times Y$, then for all $\epsilon > 0$,

$$\text{Prob}_{\mathbf{z} \in Z^m} \left\{ \left| \frac{1}{m} \sum_{i=1}^{m} \zeta(z_i) - E[\zeta] \right| \geq \epsilon \right\} \leq 2e^{\frac{-m\epsilon^2}{2\left(\sigma^2(\zeta) + \frac{1}{3}M\epsilon\right)}} \tag{41}$$

The next Theorem makes use of the above result and the compactness of $\mathcal{H}$ to derive a uniform estimate on the defect. We define the covering number $\mathcal{N}(S,s)$ for a metric space S and $s > 0$ to be the minimum number of balls with radius s that covers S. If S is compact, then $\mathcal{N}(S,s)$ is finite.

Theorem 7 ([1], Chapter 1, Theorem B): Let $\mathcal{H} \subset \mathcal{C}_X$ be compact, and assume that $\exists M > 0 : \forall f \in \mathcal{H}, |f(x) - y| \leq M$ almost everywhere. Then, $\forall \epsilon > 0$,

$$\text{Prob}_{\mathbf{z} \in Z^m} \left\{ \sup_{f \in \mathcal{H}} |L_{\mathbf{z}}(f)| \leq \epsilon \right\} \geq 1 - \delta(\epsilon, m, M) \tag{42}$$

where

$$\delta(\epsilon, m, M) = \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{8M}\right) 2e^{\frac{-m\epsilon^2}{4\left(2\sigma^2 + \frac{1}{3}M^2\epsilon\right)}} \tag{43}$$

and,

$$\sigma^2 = \sup_{f \in \mathcal{H}} \sigma^2(f^2) \tag{44}$$

Before proving the result, we state the following observations, which hold consistent with our intuition on the problem.

- $\delta(\epsilon, m, M)$ is inversely proportional to m. As more samples are available, better guarantees can be given on the error.
- $\delta(\epsilon, m, M)$ increases as both σ^2, M increase. This captures the fact that a *better* guarantee can be given on the error for spaces of functions with lower maximum variation.
- As the dimension of $\mathcal{H}$ increases, so does $\mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{8M}\right)$ and consequently, $\delta(\epsilon, m, M)$. Also, σ^2 - whose contribution is in the same direction - increases.

where the last comment introduces what is known as the bias variance tradeoff. That is, fixing the number of sample points, as $\mathcal{H}$ becomes more *complex*, while the approximation error decreases, as the estimate $f_{\mathcal{H}}$ becomes more accurate, but, the sample error increases, as the fixed number of samples becomes insufficient to learn the best estimate over $\mathcal{H}$. Hence, the problem of choosing the optimal size of $\mathcal{H}$ is, in general, non trivial. Section VI sheds more light on this issue. We now proceed with the proof of Theorem 7. We first prove the following lemma, that is needed in our proof.

Lemma 1: If $|f_j(x) - y| \leq M$ on a set $U \subset Z$ of full measure for $j = 1, 2$, then for $\mathbf{z} \in U^m$,

$$|L_{\mathbf{z}}(f_1) - L_{\mathbf{z}}(f_2)| \leq 4M \|f_1 - f_2\|_\infty \tag{45}$$

Proof: we first provide bounds on $|\mathcal{E}(f_1) - \mathcal{E}(f_2)|$ and $|\mathcal{E}_{\mathbf{z}}(f_1) - \mathcal{E}_{\mathbf{z}}(f_2)|$, then use the triangle inequality to derive the desired bound.

$$\begin{aligned}
|\mathcal{E}(f_1) - \mathcal{E}(f_2)| &= \left| \int (f_1(x) - y)^2 - (f_2(x) - y)^2 \right| && (46)\\
&= \left| \int (f_1(x) - f_2(x)) (f_1(x) + f_2(x) - 2y) \right| && (47)\\
&\leq \|f_1 - f_2\|_\infty \left| \int (f_1(x) + f_2(x) - 2y) \right| && (48)\\
&\leq \|f_1 - f_2\|_\infty \int |(f_1(x) - y) + (f_2(x) - y)| && (49)\\
&\leq \|f_1 - f_2\|_\infty \left(\int |(f_1(x) - y)| + \int |(f_2(x) - y)| \right) && (50)\\
&\leq \|f_1 - f_2\|_\infty 2M && (51)
\end{aligned}$$

where the integration is done with respect to ρ. Also,

$$\begin{aligned}
|\mathcal{E}_{\mathbf{z}}(f_1) - \mathcal{E}_{\mathbf{z}}(f_2)| &= \frac{1}{m} \left| \sum_{i=1}^{m} (f_1(x_i) - f_2(x_i))(f_1(x_i) + f_2(x_i) - 2y_i) \right| && (52)\\
&\leq \|f_1 - f_2\|_\infty \frac{1}{m} \sum_{i=1}^{m} |(f_1(x_i) - y_i)| + |(f_2(x_i) - y_i)| && (53)\\
&\leq \|f_1 - f_2\|_\infty 2M && (54)
\end{aligned}$$

It follows by the triangle inequality that,

$$\begin{aligned} |L_{\mathbf{z}}(f_1) - L_{\mathbf{z}}(f_2)| &= |\mathcal{E}(f_1) - \mathcal{E}_{\mathbf{z}}(f_1) + \mathcal{E}(f_2) - \mathcal{E}_{\mathbf{z}}(f_2)| && (55)\\ &\leq \|f_1 - f_2\|_\infty 4M && (56)\end{aligned}$$

■

Proof of Theorem 7: Consider a covering of $\mathcal{H}$ by $l = \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{8M}\right)$ balls $\{b_1, b_2, \ldots, b_l\}$, each of radius $\frac{\epsilon}{8M}$. We first bound the probability that the supremum of $|L_{\mathbf{z}}(f)|, \forall f \in b_j, j \in [l]$ in each ball, then use the union bound of probability to give the uniform bound on $\mathcal{H}$. Let $\{f_1, f_2, \ldots, f_l\}$ be the centers of $\{b_1, b_2, \ldots, b_l\}$, with respect to order. As $|f(x) - y| \leq M$ almost everywhere, we let $U \subset Z$ be a set of full measure on which this relation holds. We know from Lemma 1 that $\forall \mathbf{z} \in U^m, \forall f \in b_j, j \in [l]$,

$$\begin{aligned} |L_{\mathbf{z}}(f) - L_{\mathbf{z}}(f_j)| &\leq 4M\|f - f_j\|_\infty && (57)\\ &\leq \frac{\epsilon}{2} && (58)\end{aligned}$$

It follows that,

$$\sup_{f \in b_j} |L_{\mathbf{z}}(f)| \geq \epsilon \Rightarrow |L_{\mathbf{z}}(f_j)| \geq \frac{\epsilon}{2} \quad (59)$$

Hence, a bound that is uniform on all functions in b_j can be derived from a bound that applies on the center function f_j. To summarize, Bernstein's inequality leads to pointwise bounds on the defect, then Lemma 1 uses this bound to find a uniform bound within each ball b_j. Then, the compactness of $\mathcal{H}$ and the union bound of probability lead to a uniform bound on the defect. More precisely, let $j \in [l]$, then,

$$\begin{aligned} \text{Prob}_{z \in Z^m}\left\{\sup_{f \in \mathcal{H}} |L_{\mathbf{z}}(f)| \geq \epsilon\right\} &\leq \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{8M}\right) \text{Prob}_{z \in Z^m}\left\{\sup_{f \in b_j} |L_{\mathbf{z}}(f)| \geq \epsilon\right\} && (60)\\ &\leq \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{8M}\right) \text{Prob}_{z \in Z^m}\left\{|L_{\mathbf{z}}(f_j)| \geq \frac{\epsilon}{2}\right\} && (61)\\ &\leq \mathcal{N}\left(\mathcal{H}, \frac{\epsilon}{8M}\right) 2e^{\frac{-m\epsilon^2}{4\left(2\sigma^2 + \frac{1}{3}M^2\epsilon\right)}} && (62)\end{aligned}$$

■

B. Bounding the Approximation Error

For the same setting considered in deriving a bound on the sample error, [[1], Chapter 2] provides a result that guarantees a bound on the approximation error. i.e., the part of the error that results from the choice of the hypothesis space. A natural question that arises before considering the analysis, is why do not we consider the whole space, and restrict our attention to a smaller hypothesis subspace. One obvious reason is due to the complexity of the optimization algorithm that finds the best estimate. However, even if the best empirical estimate is assumed to be available for any choice of the hypothesis, as we observed above, the sample error increases as the size of the hypothesis space does. The following example considers the case where functions $f : X \to Y$ can have arbitrary high frequencies, hence their empirical estimation has high sensitivity to errors. There, a good candidate for the hypothesis subspace is the space of such functions with a band limit.

Let X be the n-dimensional torus. For each $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbf{Z}^n$, let $\phi_\alpha(x) = (2\pi)^{-n/2} e^{i(\alpha x)}, \forall x \in X$, where $i = \sqrt{-1}$. Let μ be the Lebesgue measure defined on $\mathbf{R}^n$, $\mathcal{L}^2_\mu(X)$ is the space of square integrable functions on X with respect to μ. Note that the set $\{\phi_\alpha\}_{\alpha \in \mathbf{Z}^n}$ is a Hammel basis of $\mathcal{L}^2_\mu(X)$ with respect to the L_2 inner product on μ. Thus, $\forall f \in \mathcal{L}^2_\mu(X)$,

$$f = \sum_{\alpha \in \mathbf{Z}^n} c_\alpha \phi_\alpha \quad (63)$$

Figure 3 illustrates the fact that if α is allowed to be arbitrarily large, then each of the basis functions ϕ_α can have high frequency components, which the samples may not capture due to physical limitations of the measuring devices. Note, however that such a phenomenon is not captured in the above model, as we assume that the same probability measure ρ used for measuring the error, is the one governing the sampling process.

For the above discussed reason, the search for the best estimate is restricted over the hypothesis space $\mathcal{H}_N$ spanned by $\{\phi_\alpha\}_{\|\alpha\|^2 \leq B}$, where $N = N(B)$ is the number of integer lattice points in the ball of radius B of $\mathbf{R}^n$. f_ρ is assumed to be bounded, hence lies in both $\mathcal{L}^2_\mu(X)$ and $\mathcal{L}^2_\rho(X)$. We here note the following about the upper bound stated in [[1], Chapter 2, Theorem 1].

- As expected, the upper bound is proportional to $\frac{Vol(X)}{N}$, where $Vol(X)$ denotes the volume of X. In other words, As we consider a larger hypothesis space, we can guarantee a lower approximation error.
- The bound is proportional to $D_{\mu\rho}$, which captures the distortion introduced by ρ to the Lebesgue measure μ, measured by the norm of the identity function from $\mathcal{L}^2_\mu(X)$ to $\mathcal{L}^2_\rho(X)$. It is not clear to us, whether this is fundamental, or just a

result of the bounding proof that first finds a bound on the distance between f_ρ and $f_{\mathcal{H}}$ with respect to μ, then uses the fact,

$$\|f_\rho - f_{\mathcal{H}}\|_\rho \leq D_{\mu\rho}\|f_\rho - f_{\mathcal{H}}\|_\mu \tag{64}$$

where $\|f\|_\nu$ is the L_2 norm of f with respect to the measure ν.

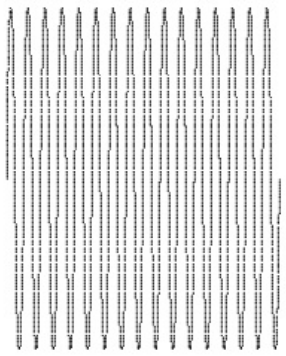


Fig. 3: ϕ_α when α is large. $n = 1$.

Before concluding this section, it is important to recall that all the insights driven here, stem from upper bound analysis, and may not reflect features of both the sample and approximation errors. Therefore, finding lower bounds on the error is essential to asserting those observations. It is worth noting that in the context of neural networks, it is mentioned in [2] that such lower bounds *do not seem to exist in the literature*.

VI. Choosing the Hypothesis Space

Consider the analysis in the previous section. Fix the probabilistic confidence level on the sample error δ, and derive a function $\epsilon(N)$ such that,

$$\text{Prob}_{\mathbf{z}\in Z^m}\left\{\mathcal{E}_{\mathcal{H}_N}(\hat{f}_{\mathbf{z},\mathcal{H}_N}) \leq \epsilon(N)\right\} \geq 1 - \delta \tag{65}$$

Also, from [[1], Chapter 2, Theorem 1], one can derive a function $A(N)$ such that,

$$\mathcal{E}(f_{\mathcal{H}_N}) \leq A(N) \tag{66}$$

where, as N increases, $\epsilon(N)$ increases and $A(N)$ decreases. Hence, in this context, it is desired to select the value of N that minimizes the sum $\epsilon(N) + A(N)$.

To provide further evidence of the phenomenon of the bias variance tradeoff in selecting the right dimension of the hypothesis space, we mention here - in brief - the example considered in [[2], Chapter 2]. There, the error criterion, as well as the choice of Y, and the probability measure ρ, are the same as above in Section IV, $X = \mathbf{R}^k$. The hypothesis space consists of the class of Radial Basis function neural networks of N nodes. i.e.,

$$f(\mathbf{x}) = \sum_{i=1}^{N} \beta_i G\left(\frac{\|\mathbf{x} - \mathbf{t}_i\|}{\alpha_i}\right) \tag{67}$$

where G is a given Gaussian basis function, and β_i, $\mathbf{t}_i$, α_i, are parameters, over which we search during the learning process.

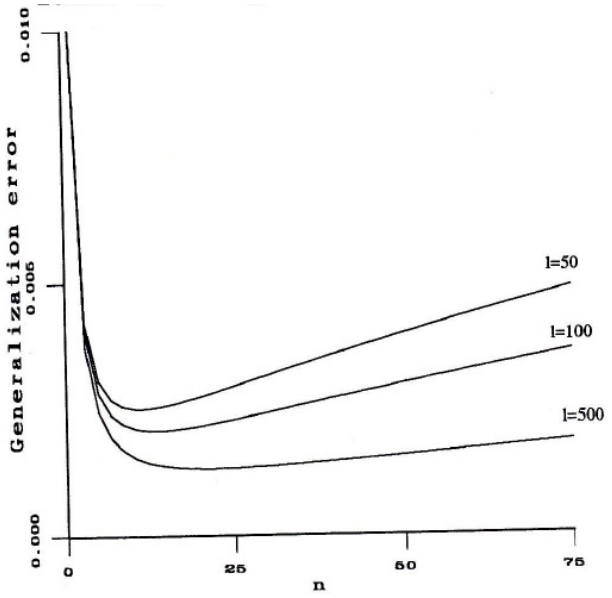
The space $\mathcal{F}$ considered is the *Liouville Space* of order n. For precise definition of the space, refer to [[2], P. 35]. Here, we note the following.

- The definition of the space $\mathcal{F}$ imposes regularity properties on the measure ρ.
- Here, in a rather unnatural way of posing the problem, the class of functions $\mathcal{F}$ was chosen such that the designated choice of the structure of the hypothesis space (RBF neural networks) becomes a good candidate.
- A bound similar to the one limiting the variations of the considered functions in the condition of Theorem 7 applies both on the space $\mathcal{F}$ and the approximating class. For the latter, it is assumed that the following holds.

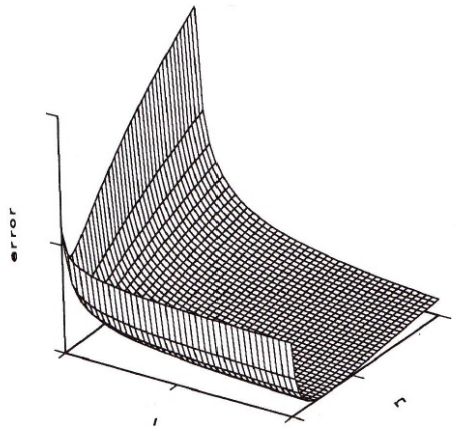
$$\sum_{i=1}^{N} |\beta_i| \leq M \tag{68}$$

We note that this condition, as in the analysis above, is needed only for bounding the sample error.

Figure 4 plots the upper bound on error $\mathcal{E}(\hat{f}_{\mathbf{z},\mathcal{H}})$ for different values of the number of sample points and the number of nodes in the network.



(a) error upper bound versus number of nodes for different values of the cardinality of the sample vector



(b) 3-dimensional figure showing bound on the error versus both the number of nodes and the number of sample points

Fig. 4: Figure showing how the upper bound on error $\mathcal{E}(\hat{f}_{\mathbf{z},\mathcal{H}})$ changes with the number of samples l and the number of nodes in the network n. As l increases, the minimum error decreases. Fixing l, the choice of the optimal n is non trivial.

In both examples mentioned above, the choice of the hypothesis space is made based on the number of sample points. However, the option of tailoring the choice of $\mathcal{H}$ based on the actual realization of $\mathbf{z}$ was not considered. One principle related to the latter approach is that of *structural risk minimization* introduced by Vapnick and Chervonenkis. In that context, a problem of similar flavor to that of choosing a hypothesis space that is rather complex with respect to the number of sample points (e.g. n too large in Figure 4), is that of *over fitting*. i.e., tailoring the choice of $\mathcal{H}$ and consequently $\hat{f}_{\mathbf{z},\mathcal{H}}$, such that the empirical estimate poorly generalizes the behavior of the true function f_ρ at points that are not sampled. Regularization techniques - e.g. Tikhonov Regularization (see for example [[3], Chapter 3] are used to relax the complexity of the empirical estimate. Hence, two factors are considered in choosing the hypothesis space.

- One that rewards the best fit to the data
- Another factor that rewards less complex spaces

One example of that is to consider a cost function $C(f)$ for $f \in \mathcal{F}$ such that.

$$C(f) = \mathcal{E}_{\mathbf{z}}(f) + \lambda\phi(f) \tag{69}$$

where λ is a regularization parameter, $\phi(f)$ is a function that is proportional to the *smoothness* of function f.

Finally, it is worth mentioning that Tikhonov regularization, as applied to linear operators [3], can be applied to the solution introduced in Section III to reduce the error sensitivity due to a large condition number of the matrix A.

VII. Future Research

Before concluding this article, we suggest a direction for future research, that as we hope, may shed insights on novel features and solutions of the mathematical problem of learning by examples

A. Over Fitting with Noiseless Data

In the analysis above, the empirical estimate $\hat{f}_{\mathbf{z},\mathcal{H}}$ minimized the empirical estimate $\mathcal{E}_{\mathbf{z}}(f)$. Then, to combat over fitting, we mentioned the regularized solution in Equation (69) that takes into account the complexity of the estimate. However, to the best of our knowledge, the problem of over fitting has been commonly considered as a result of the possibility that the samples does not faithfully capture the true function f_ρ. We observe that besides noise, the *tolerance* of the probabilistic error criterion is one other reason for choosing a hypothesis that does not fit the sampled data exactly.

More precisely, consider a setting where we know that $\forall i \in [m], f_{opt}(x_i) = y_i$. In other words, samples are only drawn from the set of points $S_{opt} = \{(x_i, y_i) : f_{opt}(x_i) = y_i\}$. Figure 5 depicts a case, where, for a given hypothesis class which includes a function that fits the data exactly, the best estimate *ignores* specific samples to minimize the generalization error. It is important to note that in the depicted example, we know that *most* points in S_{opt} lie *close* to a line. This information can be available, in the above model of Section IV, through partial knowledge of ρ. We hope to analyze settings of practical relevance under which this phenomenon holds.

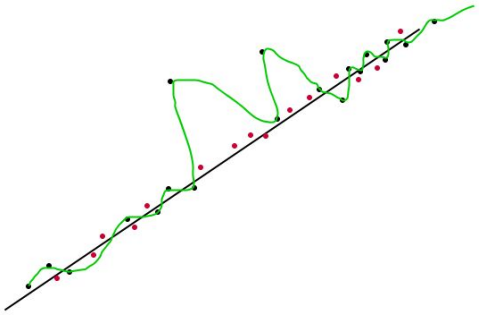


Fig. 5: Figure illustrating how over fitting can happen with noiseless data. The black dots represent samples. The green curve depicts a hypothesis that fits the data exactly. The black line depicts another hypothesis, that is simpler, yet less accurate, with respect to empirical data. The red dots represent points in S_{opt} that does not appear in the sample vector $\mathbf{z}$.

VIII. Conclusion

In this article, we presented a basic analysis for the problem of learning from examples. We first presented a linear vector space formulation, in which, the problem reduces to that of finding the least square solution of a linear operator defined by basis functions of the hypothesis space and the sample points . Then, we identified the different components of the error, namely, the sample and approximation errors. In particular, this decomposition was shown to be useful in deriving probabilistic upper bounds on the error, as well as splitting the design of a solution to a part that selects the hypothesis space, and another that finds the best estimate within that subspace, based on the empirical data. We then considered the problem of choosing the hypothesis space in light of the examples given in [1] and [2]. Finally, we suggested, as a future direction of research, considering the problem of *noiseless over fitting*.

References

[1] F. Cucker and S. Smale, "On the Mathematical Foundations of Learning," *Bulletin of the American Mathematical Study*, vol. 39, pp. 1–49, October 2001.
[2] P. Niyogi, "The Informational Complexity of Learning: Perspectives on Neural Networks and Generative Grammar,"
[3] Y. Bresler, S. Basu and C. Couvreur "Hilbert Spaces and Least Squares Methods for Signal Processing," *ECE513 Notes*, University of Illinois at Urbana Champaign, 2009.