SYSTEM IDENTIFICATION WITH UNBOUNDED LOSS FUNCTIONS UNDER ALGORITHMIC DEFICIENCY

Majid Fozunbal, Mat C. Hans†, and Ronald W. Schafer

Hewlett-Packard Laboratories
Palo Alto, CA 94304, USA

ABSTRACT

We describe and analyze a comprehensive learning model to address issues such as consistency, convergence rate, and sample complexity in the general context of system identification. The learning model is based on unbounded loss functions, and it incorporates a measure of algorithmic deficiency. We define and use a novel formulation of algorithmic solution that is an extension of the empirical risk minimization method in the sense that it uses a generic notion of side information as opposed to the commonly used input/output observation of a system. Sufficient conditions for consistency as well as closed form expressions for exponential convergence rate and sample complexity of the identification algorithm are derived.

1. INTRODUCTION

Analytical results and arguments developed in the area of learning theory are quite insightful in treating other statistical inference problems [1]. Potentially, this includes many applications in signal processing and communications such as hypothesis testing, parameter estimation, pattern recognition, channel estimation, system identification, filtering, prediction, and echo control.

Learnability is a property that is attributed to a collection of *loss functions* associated with a *probabilistic model* defined by a collection of probability distributions. The main challenge in verification of learnability for a given collection of loss functions is to prove analytical properties such as uniform convergence in probability or almost sure uniform convergence of a sequence of empirical *risk* functions. The *true risk function* is defined as the expected loss over the side information governed by a probability measure belonging to the given probabilistic model. On the other hand, the empirical risk functions are obtained by sample averaging the loss functions over the given side information, which is the basis of the most common learning method known as the empirical risk minimization (ERM) method [1].

The existence of a rich literature in learning theory and empirical processes is a great asset in dealing with similar problems in statistical signal processing. However, certain considerations need to be applied to the existing learning models to accommodate the requirements in such applications. First, the model used in signal processing applications should incorporate algorithmic deficiencies due to finite precision or finite computational resources. Second, many signal processing applications use *unbounded loss functions* such as regression function, mean square error, and Kullback-Liebler distance. While many publications in learning theory address the case of *bounded loss functions*, the case of unbounded loss functions has not attracted much attention, yet.

Motivated by earlier efforts [2], [3], we address a more comprehensive learning model for system identification. In Section 2, instead of starting from the probably approximately correct (PAC) learning model, we formulate the problem by establishing a notion of *solution operator* for system identification. We then proceed to define a generic notion of side information and algorithmic solution for this model which is based on ERM and incorporates algorithmic deficiency. Doing so, we aim to provide a more intuitive approach in articulating a broader range of applications in a learning-theoretic framework. In Section 3, we provide sufficient conditions for learnability, and we state and prove results on the rate of convergence. In Section 4, we address a certain class of unbounded loss functions with finite *VC-dimension* [1] and describe the sample complexity in learning with these loss functions. Section 5 gives some concluding remarks and future research directions.

2. MATHEMATICAL FRAMEWORK

Let $\mathcal{H} \subseteq \mathbb{R}^{N_s}$ be a compact set denoting the *state space* of a system of dimension N_s. As a common signal processing example, the state space might contain all possible acoustic coupling paths between a loadspeaker and microphone in a room where people can move around freely. Suppose the system is in an unknown state $h \in \mathcal{H}$, and we are given some side information to identify/estimate the state of the system in light of certain objective criteria. In the echo control example, side information would be the pair of speaker and microphone signals, and the objective is to find an estimate for the echo path to minimize a mean square error criterion. Let $(\Omega, \mathscr{F})$ be a measurable space and let $\mathscr{P}$ be a collection of

†Now with Motorola Laboratories, Schaumburg, IL 60196, USA.

probability measures, where for every $P \in \mathscr{P}$, $(\Omega, \mathscr{F}, P)$ is a *probability space*. For the sake of generality, we use a compact set $\mathcal{G} \subseteq \mathbb{R}^{N_\mathrm{t}}$ as the *target space* for identification, where $N_\mathrm{t} \leq N_\mathrm{s}$.

2.1. Solution Operator

To establish a comprehensive mathematical framework, we first define a notion of *problem* and *solution operator* such that given a *problem element*, one can find, at least in theory, a solution for the problem element to within the desired accuracy. In the echo control example, the problem element is the response of the unknown echo path, and the solution would be an estimate of the impulse response within some desired accuracy. We call an operator $\mathrm{S}\colon \mathcal{H} \times \mathscr{P} \times \mathbb{R}^+ \to 2^{\mathcal{G}}$ a solution operator provided that for every given $h \in \mathcal{H}$, $P \in \mathscr{P}$, and accuracy $\varepsilon \in \mathbb{R}^+$, the nonempty set $\mathrm{S}_P(h, \varepsilon) \subseteq \mathcal{G}$ denotes the *ε-approximate P-solution set* for h and $\mathrm{S}_P(h, \varepsilon_1) \subseteq \mathrm{S}_P(h, \varepsilon_2)$ if $\varepsilon_1 \leq \varepsilon_2$.

Let $\mathcal{U}$ be a subset of a finite dimensional Euclidean space called the *uncertainty space* and let $\{U_n\}$ be an independent, identically distributed (i.i.d.) stochastic process such that at each time instance n, $U_n\colon \Omega \to \mathcal{U}$ models the uncertain elements that are present in system identification. Let $\mathrm{L}\colon \mathcal{H} \times \mathcal{G} \times \mathcal{U} \to \mathbb{R}^+$, be a cost function describing the cost of approximating an element $h \in \mathcal{H}$ with an element $g \in \mathcal{G}$ when the uncertain element is $u \in \mathcal{U}$. The *risk* function is defined as the expected loss of such approximation

$$J_P(h, g) = \int \mathrm{L}(h, g, u) dP. \tag{1}$$

We define the solution operator such that

$$\mathrm{S}_P(h, \varepsilon) = \left\{ g \in \mathcal{G} : J_P(h, g) \leq \inf_{g \in \mathcal{G}} J_P(h, g) + \varepsilon \right\} \tag{2}$$

describes the ε-approximation P-solution set for h. Assuming analytical continuity of the risk function over $\mathcal{G}$, one can verify that (2) satisfies our definition of solution operator.

2.2. Algorithmic Solution

Most learning models are based on processing random side information. That is, some random side information is given about an unknown system state, and we are required to identify the unknown state with a certain desired accuracy. This is a challenging problem due to uncertainties that occur because of insufficient or corrupted side information.

Let $\mathcal{O}$ be a subset of a finite dimensional Euclidean space called the *observation space*. Side information is generated by a sequence of mappings $\boldsymbol{I} = (\mathrm{I}_n)_{n \in \mathbb{N}}$ such that for each n, $\mathrm{I}_n\colon \mathcal{H} \times \mathcal{U}^n \to \mathcal{O}^n$ called an *information operator of order n*. We assume that I_n is formed as a stationary extension of a function $I\colon \mathcal{H} \times \mathcal{U} \to \mathcal{O}$ called the *information function*. For an underlying pair (h, u^n), I_n generates a multisample $o^n = (o_i)_{i=1}^n \in \mathcal{O}^n$ as side information. The objective is to find a sequence of mappings $\boldsymbol{A} = \{\mathrm{A}_n\}$ such that $\mathrm{A}_n\colon \mathcal{O}^n \times \mathbb{R}^+ \to 2^{\mathcal{G}}$ where for given $o^n \in \mathcal{O}^n$, $o^n = (o_i)_{i=1}^n$, and an algorithmic deficiency $\delta \in \mathbb{R}^+$, $\mathrm{A}_n(o^n, \delta) \subseteq \mathrm{S}_P(h, \varepsilon)$ for arbitrarily high confidence as n becomes sufficiently large. We call $\boldsymbol{A}$ a *learning algorithm* and A_n an *algorithmic solution* of order n. In the echo control example, the side information is the sampled microphone and speaker signals and the algorithm could be the common least mean square (LMS) algorithm.

The most common approach in developing a learning algorithm is *empirical risk minimization* (ERM). To describe the ERM approach, we need to apply a restriction on the cost function as follows. We assume there exists a loss function $Q\colon \mathcal{O} \times \mathcal{G} \to \mathbb{R}^+$ such that for every h, g, and u, the cost is defined as $\mathrm{L}(h, g, u) = Q(I(h, u), g)$. Under the operation of I_n, and for a fixed but unknown $h \in \mathcal{H}$, this induces a stochastic process $\{O_n(h)\}$, called the *side information process*, where at each time instance n, $O_n(h)\colon \Omega \to \mathcal{O}$. Let P_h denote the induced distribution on O. As a result, we can describe the risk function as

$$J_P(h, g) = \int Q(g, o) dP_h. \tag{3}$$

Let $O^n(h, \omega)$ denote the beginning segment of a realization of the side information process, where for simplicity, we denote it by $o^n = (o_i)_{i=1}^n$. The *empirical risk function* is defined as

$$J^{(n)}(g) = \frac{1}{n} \sum_{i=1}^{n} Q(g, o_i). \tag{4}$$

Let $\nu^{(n)} = \inf_{g \in \mathcal{G}} J^{(n)}(g)$, the algorithmic solution of order n introduces

$$\mathrm{A}_n(o^n, \delta) = \{ g \in \mathcal{G} : J^{(n)}(g) \leq \nu^{(n)} + \delta \}, \tag{5}$$

as the empirical δ-solution set for h, where $0 \leq \delta < \varepsilon$. We define

$$\Psi_{n,P}(h, \varepsilon, \delta) \triangleq \{ \omega \in \Omega : A_n(O^n(h, \omega), \delta) \nsubseteq \mathrm{S}_P(h, \varepsilon) \} \tag{6}$$

as the event that the algorithmic solution of order n with δ-deficiency misidentifies ε-approximate solutions for h. Assuming that this event is measurable for every n, we would like to investigate the behavior of

$$\sup_{P \in \mathscr{P}} \sup_{h \in \mathcal{H}} P(\Psi_{n,P}(h, \varepsilon, \delta)) \tag{7}$$

with respect to n. Equation (7) determines the worst case probability of misidentification whose convergence to zero means that the algorithm can identify any unknown state with accuracy ε. For example, in the echo control problem, convergence of (7) implies that the adaptive algorithm will attain the desired level of accuracy for any underlying echo path and and any governing probability distribution belonging to the probabilistic model.

3. CONVERGENCE AND ASYMPTOTIC BEHAVIOR

To prove consistency and convergence in probability, and to investigate the asymptotic behavior of (7), we first relate (7)

to the uniform convergence arguments known in the literature of learning theory [1], [4].

3.1. Uniform Convergence

To study convergence property for the algorithm, let

$$\Phi_{n,P}(h,\epsilon) \triangleq \Big\{\omega \in \Omega : \sup_{g\in\mathcal{G}} \Big| J_P(h,g) - J^{(n)}(g)\Big| \geq \epsilon\Big\}. \quad (8)$$

We now state and prove the following lemma.

Lemma 3.1. *For every $0 \leq \delta < \varepsilon$, the following inequality holds true*

$$P(\Psi_{n,P}(h,\varepsilon,\delta)) \leq P(\Phi_{n,P}(h,(\varepsilon-\delta)/2)). \quad (9)$$

Proof. Let $J_P^o(h) = \inf_{g\in\mathcal{G}} J_P(h,g)$. One can verify that

$$\Psi_{n,P}(h,\varepsilon,\delta) = \{\omega \in \Omega : \exists g \in \mathcal{G},\ J^{(n)}(g) \leq \nu^{(n)} + \delta, \\ J_P(h,g) > J_P^o(h) + \varepsilon\}.$$

Let $\epsilon = (\varepsilon-\delta)/2$. We define

$$\Delta_n(\epsilon) \triangleq \Big\{\omega \in \Omega : J_P^o(h) - \nu^{(n)} \geq -\epsilon\Big\}$$

and its complement $\Delta_n^c(\epsilon) = \Omega\backslash\Delta_n(\epsilon)$. One can verify that

$$\Psi_{n,P}(h,\varepsilon,\delta) \subseteq (\Psi_{n,P}(h,\varepsilon,\delta) \cap \Delta_n(\epsilon)) \cup \Delta_n^c(\epsilon).$$

Consider the event $\Psi_{n,P}(h,\varepsilon,\delta)\cap\Delta_n(\epsilon)$ and let ω be an outcome belonging to this event. By definition, there exists a $g \in \mathcal{G}$ that satisfies the conditions of $\Psi_{n,P}(h,\varepsilon,\delta)$. We can verify that for this outcome ω, the inequality

$$J_P(h,g) - J^{(n)}(g) \geq \varepsilon - \delta + J_P^o(h) - \nu^{(n)} \geq \frac{\varepsilon-\delta}{2}$$

holds true. As a result, for every outcome $\omega \in \Psi_{n,P}(h,\varepsilon,\delta)\cap \Delta_n(\epsilon)$, we have

$$\sup_{g\in\mathcal{G}}\big(J_P(h,g) - J^{(n)}(g)\big) \geq \epsilon. \quad (10)$$

Now consider an outcome $\omega \in \Delta_n^c(\epsilon)$ and note that for this outcome, we have

$$v^{(n)} - J_P^o(h) > \epsilon.$$

By continuity of $J_P(h,g)$ and compactness of $\mathcal{G}$, there exists an optimal $g^o \in \mathcal{G}$ such that $J_P^o(h) = J_P(h,g^o)$. By definition, we have $J^{(n)}(g^o) \geq \nu^{(n)}$. This means for very outcome $\omega \in \Delta_n^c(\epsilon)$, we have $J^{(n)}(g^o) - J_P(h,g^o) > \epsilon$ that implies

$$\sup_{g\in\mathcal{G}}\big(J^{(n)}(g) - J_P(h,g)\big) \geq \epsilon. \quad (11)$$

By (10) and (11), we can conclude the assertion. □

3.2. Relative Uniform Convergence

Since we are interested in unbounded loss functions, we need to apply some restriction on the set of probability measures that is assumed in modeling the problem. For this purpose, we first assume that there exists a dominating function $M : \mathcal{O} \to \mathbb{R}^+$ such that for every $o \in \mathcal{O}$,

$$\sup_{g\in\mathcal{G}} Q(o,g) \leq M(o) < \infty. \quad (12)$$

Now, using a parameter $\rho > 0$, we index and define the desired collection of probability measures as follows

$$\mathscr{P}_\rho = \Big\{P : \Big[\int M^2 dP\Big]^{1/2} \leq \rho\Big\}. \quad (13)$$

Hence, in the rest of the paper, we assume (13) in definition (7). Let us define

$$K_P(h,g) \triangleq \Big[\int Q^2 dP\Big]^{1/2},$$

and let

$$\Upsilon_{n,P}(h,\epsilon) \triangleq \left\{\omega \in \Omega : \sup_{g\in\mathcal{G}} \frac{\big|J_P(h,g) - J^{(n)}(g)\big|}{K_P(h,g)} > \epsilon/\rho\right\}.$$

The following proposition holds.

Proposition 3.1. *For every $h \in \mathcal{H}$ and $P \in \mathscr{P}_\rho$,*

$$P(\Phi_{n,P}(h,(\varepsilon-\delta)/2)) \leq P(\Upsilon_{n,P}(h,(\varepsilon-\delta)/2)).$$

Proof. The proof of this proposition follows by the observation that $K_P(h,g) \leq \rho$ for all $h \in \mathcal{H}$ and $g \in \mathcal{G}$. □

The importance of Proposition (3.1) is that it enables us to use the results of [1] in studying the convergence rate.

3.3. Exponential Convergence

Let $H_Q(n)$ denote the *growth function* [1, p.190] of the class of loss functions $\{Q(\cdot,g)\}_{g\in\mathcal{G}}$ for n observations.The following theorem provides an upper-bound on (7).

Theorem 3.1. *Given δ, ε, and ρ, such that $0 < \frac{\varepsilon-\delta}{2\rho} \leq 1$, we have*

$$\sup_{P\in\mathscr{P}_{\rho,\rho}} \sup_{h\in\mathcal{H}} P(\Psi_{n,P}(h,\varepsilon,\delta)) \\ \leq 8\exp\left\{\Big(\frac{H_Q(2n)}{n} - \frac{(\varepsilon-\delta)^2}{2^4\rho^2\xi^2}\Big)n\right\}, \quad (14)$$

where $1 \leq \xi \leq \sqrt{1 - \ln\frac{\varepsilon-\delta}{2\rho}}$ is the solution of

$$\xi = \sqrt{1 - \frac{\ln\frac{\varepsilon-\delta}{2\rho\xi}}{2}}. \quad (15)$$

Proof. To prove the assertion, we use the results of Lemma 3.1 and Proposition (3.1). By applying the results of [1, Thm. 5.4, p. 200], we conclude that for every $h \in \mathcal{H}$ and $P \in \mathscr{P}_\rho$ the above inequality holds. Taking the supremum, we conclude the assertion. To obtain the upper-bound on ξ, we simply use the inequality $\ln\xi \leq \xi - 1$ and convert the equation into a quadratic problem. □

Theorem 3.1 describes a sufficient condition for uniform convergence of the identification algorithm. Recalling the echo control example, Theorem 3.1 describes conditions that guarantee the convergence of the echo control algorithm within the desired level of accuracy.

Equation (14) demonstrates an upper bound on the worst case probability of misidentification whose convergence to zero as n grows implies consistent successful learning. Using (14), we obtain a lower bound on the *exponential convergence rate* of (7) as

$$\alpha(\varepsilon,\delta,\rho)=\liminf_{n\to\infty}\Big(\frac{(\varepsilon-\delta)^2}{2^4\rho^2\xi^2}-\frac{H_Q(2n)}{n}\Big). \qquad (16)$$

The second term of the right hand side of the equation is dependent on the growth function of the collection of loss functions. Clearly, to have an algorithm that can estimate the state of the system with any desired level of accuracy, it is sufficient that the growth function of the collection of loss functions satisfies

$$\limsup_{n\to\infty}\frac{H_Q(2n)}{n}=0. \qquad (17)$$

Intuitively, (17) means that to have a consistent learning algorithm, it suffices that as the order of information operator increases, the uncertainty about the unknown system state decreases with a faster rate. The problem of finding the growth function of a class of loss functions is not an straightforward problem and usually growth functions are upper bounded by closed-form expressions using the notion of VC*-dimension* and its generalizations [1], [5]. Provided that (17) holds for the class of loss functions, the convergence rate of the learning algorithm is specified by

$$\alpha(\varepsilon,\delta,\rho)=\frac{(\varepsilon-\delta)^2}{2^4\rho^2\xi^2}, \qquad (18)$$

where ξ is obtained from (15). Equation (18) simply expresses how the level of accuracy, the algorithmic error, and certain statistics of the collection of probability measures affect the exponential convergence rate.

4. AFFINE-BASED LOSS FUNCTIONS

While (14) indicates the learning consistency for any class of cost functions that satisfy (12) and (17), to further explore some details about the growth function, we introduce a certain type of loss function that is parameterized by affine combinations of the given class of objects.

Lemma 4.1. *Suppose $Q=T\circ R$ where $R\colon \mathcal{O}\times\mathcal{G}\to\mathbb{R}$ is affine on $\mathcal{G}$, and $T\colon \mathbb{R}\to\mathbb{R}^+$ is continuous and monotonic. Then, the growth function of the class of loss functions is upper bounded by*

$$H_Q(n)\le(N_{\rm t}+1)\left(\ln\frac{n}{N_{\rm t}+1}+1\right),$$

where $N_{\rm t}$ is the dimension of $\mathcal{G}$.

Proof. The proof follows from the results of [1, Ch. 5]. □

The importance of Lemma 4.1 is that it is applicable to common loss functions that are used in signal processing. For example, in system identification the observation space is usually the pair of input/output of the system, e.g., $\mathcal{O}=\mathcal{X}\times\mathcal{Y}$, where $\mathcal{X}=\mathbb{R}^{N_{\rm t}}$ and $\mathcal{Y}=\mathbb{R}$ denote the input space and output space of the system, respectively. Suppose $\mathcal{H}$ and $\mathcal{G}$ are the same. In this case, the squared error loss function is described by $Q(x,y,g)=|y-\langle g,x\rangle|^2$. To analyze this example, one can simply define the dominating function $M(x,y)=2|y|^2+\sup_{g\in\mathcal{G}}2\|g\|^2\|x\|^2$ and use the results discussed in Section 3.3.

Using Lemma 4.1, we now determine the *sample complexity* that is the order of information operator that guarantees a certain level of performance. Let $\beta>0$ be the worst case probability of misidentification, i.e., (7). Then, the probability of confidence in the obtained solution is larger than $1-\beta$. Using the results we have shown so far, we can obtain the following upper bound on sample complexity

$$n=\frac{2^4\rho^2\xi^2}{(\varepsilon-\delta)^2}\left(\ln\frac{8}{\beta}+(N_{\rm t}+1)\big(\ln\frac{2n}{N_{\rm t}+1}+1\big)\right), \qquad (19)$$

where $n>(N_{\rm t}+1)/2$. Equation (19) is a novel expression that demonstrates the effect of different parameters in determining the order of information operator required in approximating an element of $\mathcal{H}$ within the accuracy ε where the algorithmic deficiency is δ and confidence is no smaller than $1-\beta$. It can be seen that the accuracy ε, the algorithmic error δ, and the moment constraint of the loss function, ρ, have dramatic effects in sample complexity. For the echo control problem, (19) describes how fast the echo control algorithm reduces the echo into a desired level for exciting signals whose distributions belong to $\mathscr{P}_\rho$.

5. CONCLUSION

In this paper, we introduced a learning model for the problem of system identification, where the model is based on unbounded loss functions and it incorporates algorithmic deficiency. Convergence analysis and conditions for consistency for the algorithmic solution have been explored. It remains for future research to further explore the growth function of a class of unbounded loss functions. Moreover, it is very useful to investigate how to use such tools from learning theory to address issues in tracking time-varying systems.

6. REFERENCES

[1] Vladimir Vapnik, *Statsitical Learning Theory*, John Wiley & Sons, 1998.

[2] L. Ljung, "PAC-learning and asymptotic system identification theory," *Proceedings of the 35th IEEE Conference on Decision and Control*, pp. 2303–2307, Dec. 1996.

[3] Majid Fozunbal, Mat Hans, and Ronald W. Schafer, "A decision-making framework for acoustic echo cancelation," *Proc. IEEE Workshop on Appl. Signal Process. Audio and Acoustics*, Oct. 2005.

[4] M. Vidyasagar, *A Theory of Learning and Generalization*, Springer-Verlag, 1997.

[5] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, "Scale sensitive dimensions, uniform convergence, and learnability," *Journal of ACM*, vol. 44, pp. 616–631, 1997.