Indian Institute of Technology, Kanpur
Statistical and Algorithmic Learning Theory

Instructors: Purushottam Kar
Authors: Gurpreet Singh
Email: guggu@iitk.ac.in

SCRIBE
5

Uniform Convergence and PAC Learnability

In order to understand pointwise and uniform convergence, let us consider the following problem.

Problem Setting

Consider a finite hypothesis class $\mathcal{H} = \left\{f^1, f^2 \ldots f^m\right\}$ and a distribution $\mathcal{D}$. We need to find the best prediction function f^* which minimizes the l-risk.

Since our hypothesis space is finite, we can, without loss of generality, say that $f^* = f_1$. Then,

1. we do not want $f_2, f_3 \ldots f_m$ to perform well on S *i.e.* have high(er) empirical risk or training error.
2. we do not want f_1 to perform poorly on S *i.e.* have low(er) empirical risk.
3. we try to ensure all f_i give faithful and honest performance on S *i.e.* $\forall\, f \in \mathcal{H}, \mathrm{er}^l_D\left[\, f\,\right] \approx \mathrm{er}^l_S\left[\, f\,\right]$

Essentially, we wish to find a training sample S that for all functions in the hypothesis class, S should perform well.

We say that we need f_1 to give good performance, however we need to define what is "good". We say that S is good with respect to a function $f \in \mathcal{H}$ ($S \in \mathrm{good}_f(\epsilon)$) if

$$\left|\mathrm{er}^l_D\left[\, f\,\right] - \mathrm{er}^l_S\left[\, f\,\right]\right| \quad \leq \quad \epsilon$$

We define this more formally in the Section 2

Note. By good, we do not mean that the function should give a low error, instead, the function should give an error almost the same as it would give for the true distribution

Pointwise Convergence

Before we look at the definition of pointwise convergence, we need to formally define when a training sample is considered to be good or representative of the true data distribution.

Definition 5.1 (ϵ-representative wrt[1] a function)**.** A training sample S is said to be ϵ-representative

[1] wrt is an abbreviation for "with respect to"

of a distribution $\mathcal{D}$ with respect to a function f and a loss function l if

$$\left| \mathrm{er}^{l}_{D}\left[\, f \,\right] - \mathrm{er}^{l}_{S}\left[\, f \,\right] \right| \leq \epsilon \tag{1}$$

We can now give the formal definition of pointwise convergence.

Definition 5.2 (Pointwise Convergence)**.** A hypothesis / function class $\mathcal{H} = \left\{ f^1, f^2 \ldots \right\}$ is said to be in pointwise convergence if $\forall\, f \in \mathcal{F}$

$$\forall \epsilon > 0, \quad \lim_{n \to \infty} \mathop{\mathbb{P}}_{S \sim \mathcal{D}^n} \left[\left| \mathrm{er}^{l}_{D}\left[\, f \,\right] - \mathrm{er}^{l}_{S}\left[\, f \,\right] \right| > \epsilon \right] \quad = \quad 0 \tag{2}$$

If the hypothesis class is finite, then using Theorem 3.1, we can formulate the following observation.

Theorem 5.1. For a finite hypothesis class $\mathcal{H}$ and a loss function l, we say that $\mathcal{H}$ is in pointwise convergence if for some $\epsilon > 0$, $\delta \in (0, 1)$ and a training sample S of size n, where $n > n_{\mathcal{H}}(\epsilon, \delta)$ if $\forall\, f \in \mathcal{H}$

$$\mathop{\mathbb{P}}_{S \sim \mathcal{D}^n} \left[\left| \mathrm{er}^{l}_{D}\left[\, f \,\right] - \mathrm{er}^{l}_{S}\left[\, f \,\right] \right| > \epsilon \right] \quad \leq \quad \delta \tag{3}$$

Exercise 5.1. Prove Theorem 5.1 using Theorem 3.1 from the third scribe.

Corollary 5.1.1. If the output space $\mathcal{Y} = \{0, 1\}$ *i.e.* the problem is a binary classification one, then for any sample $S \overset{iid}{\sim} D^n$, $\forall\, f \in \mathcal{H}$

$$\mathbb{P} \left[\left| \mathrm{er}^{l}_{D}\left[\, f \,\right] - \mathrm{er}^{l}_{S}\left[\, f \,\right] \right| > \epsilon \right] \quad \leq \quad 2 \exp \left(\frac{-n\epsilon^2}{3} \right)$$

Corollary 5.1.2. If the output space $\mathcal{Y} = \{0, 1\}$ *i.e.* the problem is a binary classification one, then we can say that the minimum sampling complexity

$$n_{\mathcal{H}} \quad \leq \quad \left\lceil \frac{3 \log \left(2/\delta \right)}{\epsilon^2} \right\rceil$$

Exercise 5.2. Prove Corollary 5.1.1

Hint: Take a Bernoulli random variable and use Chernoff's Bound

Pointwise convergence seems to be the solution to the problem discussed in Section 1, however if we take a closer look, it is not a solution to our problem. Pointwise convergence assures that each function individually does not break the ϵ bound, however, if we find the bound that no function f breaks the ϵ bound, we will find the probability to be greater than $1 - |\,\mathcal{H}\,| \cdot \delta$ *i.e.*

$$\mathbb{P} \left[\, \forall\, f \in \mathcal{H},\ S \text{ is } \epsilon\text{-representative } \right] \quad \geq \quad 1 - |\,\mathcal{H}\,| \cdot \delta$$

Exercise 5.3. Prove the above bound using the concepts of pointwise convergence

This is a much weaker bound, as the size of the hypothesis space is generally large (sometimes infinite, as we will see later). Therfore, pointwise convergence is a good property for individual functions, however, we want such a property to hold for all functions collectively. Hence, we define a much stronger convergence property, called uniform convergence, discussed in the next section.

Uniform Convergence

We have stated what we desire when we choose a training sample in Section 1. As discussed in the previous section, we have concluded that pointwise convergence is a weak property and cannot be used to find a good bounds on the training sample. The solution to that is Uniform Convergence. In order to define Uniform Convergence, we must first understand for what training sample do we say that the sample is good or representative of the distribution.

Definition 5.3. Given a hypothesis space $\mathcal{H}$ and a loss fucntion l, a training sample S is called ϵ-representative of any distribution $\mathcal{D}$ if $\forall f \in \mathcal{H}$

$$\left| \mathrm{er}^{l}_{D}\,[\,f\,] - \mathrm{er}^{l}_{S}\,[\,f\,] \right| \quad \leq \quad \epsilon \tag{4}$$

Similar to pointwise converge in idea, we now define uniform convergence

Definition 5.4 (Uniform Convergence)**.** A hypothesis / function class $\mathcal{H} = \left\{ f^1, f^2 \ldots \right\}$ is said to be in uniform convergence if

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} \mathop{\mathbb{P}}_{S\sim\mathcal{D}^n} \left[\max_{f\in\mathcal{H}} \left\{ \left| \mathrm{er}^{l}_{D}\,[\,f\,] - \mathrm{er}^{l}_{S}\,[\,f\,] \right| \right\} > \epsilon \right] \quad = \quad 0 \tag{5}$$

Again, we can formulate the definition of uniform convergence to fit in our problem of hypothesis spaces.

Theorem 5.2. For a finite hypothesis class $\mathcal{H}$ and a loss function l, we say that $\mathcal{H}$ is in uniform convergence if for some $\epsilon > 0$, $\delta \in (0, 1)$ and a training sample S of size n, where $n > n_{\mathcal{H}}(\epsilon, \delta)$ if

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n} \left[\, \forall f \in \mathcal{F}, \left\{ \left| \mathrm{er}^{l}_{D}\,[\,f\,] - \mathrm{er}^{l}_{S}\,[\,f\,] \right| > \epsilon \right\} \right] \quad \leq \quad \delta \tag{6}$$

We will give the bound for $n_{\mathcal{H}}$ later in the scribe, however, note that it is still polynomial in $1/\epsilon$ and $1/\delta$

Since we define $\hat{f} = \arg\min_{f\in\mathcal{H}} \; \mathrm{er}^{l}_{S}\,[\,f\,]$, we can say

$$\mathrm{er}^{l}_{S}[\,\hat{f}\,] \quad \leq \quad \mathrm{er}^{l}_{S}\,[\,f^*\,]$$

Suppose $S \in \epsilon$-representatives[2], then

$$\begin{aligned} \mathrm{er}^{l}_{D}[\,\hat{f}\,] \quad &\leq \quad \mathrm{er}^{l}_{S}[\,\hat{f}\,] + \epsilon \\ &\leq \quad \mathrm{er}^{l}_{S}\,[\,f^*\,] + \epsilon \\ &\leq \quad \mathrm{er}^{l}_{D}\,[\,f^*\,] + 2\epsilon \end{aligned}$$

[2] We use ϵ-reprentatives to denote the set of samples that are ϵ-representative for a hypothesis class

Hence, we can say that if

$$S \in \epsilon\text{-representative} \qquad \Longrightarrow \qquad \text{er}^l_D[\,\hat{f}\,] \leq \text{er}^l_D\,[\,f^*\,] + \frac{\epsilon}{2}$$

Note that the RHS defines the condition for PAC learnability. Hence, we can formulate the following result.

Result 5.4.1. If a hypothesis class $\mathcal{H}$ is in uniform convergence for a training sample S with sample complexity $n^{\text{UC}}_{\mathcal{H}}$, then the hypothesis class $\mathcal{H}$ is PAC learnable with minimum sample complexity $n_{\mathcal{H}} \leq n^{\text{UC}}_{\mathcal{H}}$, and henceforth, the $\text{ERM}_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$

Suppose we have a training sample S for which the hypothesis class is in uniform convergence, then we also have PAC learnability for that hypothesis class. Hence, the problem discussed in Section 1 is solved with the ERM paradigm, without the fear of overfitting.

Agnostic PAC learnibility for Finite Hypothesis Classes

Similar to the method used in Scribe 3, we need to find a training sample size which allows for uniform convergence given $\epsilon > 0$ and $\delta \in (0, 1)$ in case of finite hypothesis classes.

Suppose, for a hypothesis class $\mathcal{H}$, we have a training sample S of size n. Then for some $\epsilon > 0$ and $\delta \in (0, 1)$,

$$\begin{aligned} S \in \epsilon\text{-representatives} &\implies \forall\, f \in \mathcal{H}, \; \Big|\text{er}^l_D\,[\,f\,] - \text{er}^l_S\,[\,f\,]\Big| \leq \epsilon \\ \therefore S \notin \epsilon\text{-representatives} &\implies \exists\, f \in \mathcal{H}, \; \Big|\text{er}^l_D\,[\,f\,] - \text{er}^l_S\,[\,f\,]\Big| > \epsilon \end{aligned}$$

Here, we use an identity, the proof of which is left as an exercise.

$$\text{If } \; A \implies B \; \text{ then } \; \mathbb{P}\,[\,A\,] \leq \mathbb{P}\,[\,B\,]$$

Therefore, we can say

$$\begin{aligned} \mathbb{P}\,[\,S \notin \epsilon\text{-representatives}\,] \quad &\leq \quad \mathbb{P}\Big[\,\exists\, f \in \mathcal{H}, \; \Big|\text{er}^l_D\,[\,f\,] - \text{er}^l_S\,[\,f\,]\Big| > \epsilon\Big] \\ &= \quad \mathbb{P}\left[\,\bigcup_{f \in \mathcal{H}} \Big\{\Big|\text{er}^l_D\,[\,f\,] - \text{er}^l_S\,[\,f\,]\Big| > \epsilon\Big\}\right] \\ &\leq \quad \sum_{f \in \mathcal{H}} \mathbb{P}\Big[\Big|\text{er}^l_D\,[\,f\,] - \text{er}^l_S\,[\,f\,]\Big| > \epsilon\Big] \end{aligned}$$

We can further reduce the RHS of the above inequality in terms of the size of the sample S, ϵ and the size of the hypothesis set, and hence state the condition or bound for uniform convergence. However, for this we require a concentration bound known as the *Hoeffding's Inequality*, which is as stated below

Theorem 5.3 (Hoeffding's Inequality). Let R have a sequence of n i.i.d. random variables $\{X^i\}_{i\in[\,n\,]}$ and assume that for all $i \in [\,n\,]$, $\mathbb{E}\left[\,X^i\,\right] = \mu$ and $a \leq X^i \leq b$ a.s.[3]. Then, for any $\epsilon > 0$

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i\in[\,n\,]} X^i - \mu\right| > \epsilon\right] \quad \leq \quad 2\exp\left(-\frac{2\,n\,\epsilon^2}{(b-a)^2}\right) \tag{7}$$

Remark. The Hoeffding's Inequality can be proved using a lemma, known as the *Hoeffding's Lemma*, which states for any random variable $X \in [\,a, b\,]$ a.s., such that $\mathbb{E}\,[\,X\,] = 0$, we have

$$\mathbb{E}\left[\,e^{\lambda X}\,\right] \quad \leq \quad \exp\left(\frac{\lambda^2(b-a)^2}{8}\right) \tag{8}$$

We do not give the proof of the above theorem here, but can be found here

Assuming the range of the loss function l is $[\,0, 1\,]$, then using the Hoeffding's Inequality, we can complete the bound on the probability of a training sample S not being ϵ-representative and formulate the following result, the proof of which is left as an exercise

Result 5.4.2. For a hypothesis class $\mathcal{H}$ and a loss function l, the range of which is $[\,0, 1\,]$, the hypothesis class is in uniform convergence with minimum sample complexity

$$n_{\mathcal{H}}^{\mathrm{UC}} \quad \leq \quad \left\lceil \frac{\log\left(2\,|\,\mathcal{H}\,|\,/\delta\right)}{2\epsilon^2} \right\rceil \tag{9}$$

Exercise 5.4. Find the bound on $n_{\mathcal{H}}^{\mathrm{UC}}$ if the range of the loss function is $[\,a, b\,]$

This result suggests that every hypothesis class that is finite is in uniform convergence for a training sample with size at least that of $n_{\mathcal{H}}^{\mathrm{UC}}$. Using Result 5.4.1, we can also get the following result.

Result 5.4.3. For a hypothesis class $\mathcal{H}$ and a loss function l, the range of which is $[\,0, 1\,]$, the hypothesis class is PAC learnable with minimum sample complexity

$$n_{\mathcal{H}} \quad \leq \quad n_{\mathcal{H}}^{\mathrm{UC}} \quad \leq \quad \left\lceil \frac{2\log\left(2\,|\,\mathcal{H}\,|\,/\delta\right)}{\epsilon^2} \right\rceil \tag{10}$$

[3] We say that an event E happens almost surely (abbreviated as a.s.) if the probability of E happening is 1