4 Uniform Convergence $\implies$ Generalization

Most of this section will rely on the notation introduced in Section 3.1. Our objective is to relate the generalization performance of a learning algorithm to certain properties of the problem at hand. We have already done this in the case MLE, where we characterized the behavior of the excess risk as $R(\hat{\theta}) - R(\theta_*) \approx d/n$ where d is the dimension of the features and n is the number of samples. This characterization tells us that as the number of samples increase, the excess risk decrease with a rate of n^{-1}, and as the dimension of the features increase, the excess risk also increase with a rate of d. But there were a couple of limitations of this result. First, this result was asymptotic, i.e. it only holds when $n \to \infty$. Second, the entire MLE setup assumes that we know the true parametric form of the data distribution. These are very strong assumptions which do not hold in practice.

In the sequel, our objective is modified. We do not assume that the data distribution is known anymore. We only assume that data samples are iid from some distribution. The problem we consider can be summarized as follows.

$$f_* = \operatorname*{argmin}_{f \in \mathcal{F}} \; R(f) \coloneqq \mathbb{E}[\ell((y, x), f)]$$

As before the expectation is over the true unknown distribution $(y, x) \sim p(y, x)$; thus, we cannot compute this expectation. Luckily, we can estimate this risk with a sample mean estimator (aka empirical risk). That is,

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \; \hat{R}(f) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \ell((y_i, x_i), f)$$

Notice that $\hat{R}(f)$ is an estimator to $R(f)$ and when n is large, they will be close to each other. The hope is that, their minimizers are also close, and we will show that they actually are!

Two quantities have a big impact on the generalization performance of our learning algorithm: 1- the complexity of the function class $\mathcal{F}$, and 2- the number data points used in training. We would like to characterize the behavior of the excess risk in the following way.

$$R(\hat{f}) - R(f_*) \leq \frac{\text{a func of comp of } \mathcal{F}}{\text{a func of } n}. \tag{4.1}$$

In the case of MLE, we sort of achieved this (d is not really a complexity measure of the function class but ...).

We notice that the left hand side of (4.1) is a random variable. Therefore, we need to make a probabilistic argument for this statement to make sense (e.g. almost sure, or high probability etc). We choose high probability. More formally:

$$\mathbb{P}\big(\underbrace{R(\hat{f}) - R(f_*) \geq \epsilon}_{\text{bad event}} \big) < \underbrace{\delta}_{\text{small probability}}$$

Here, ϵ and δ are ideally smaller numbers.

4.1 From excess risk to empirical process

We can decompose the excess risk in three terms.

$$R(\hat{f}) - R(f_*) = \underbrace{[R(\hat{f}) - \hat{R}(\hat{f})]}_{\text{not iid sum}} + \underbrace{[\hat{R}(\hat{f}) - \hat{R}(f_*)]}_{\leq 0} + \underbrace{[\hat{R}(f_*) - R(f_*)]}_{\text{iid sum}/n}.$$

The first term above is the main term we need heavy lifting. This is because $\hat{f}$ is a random variable, and it breaks the iid sum structure of the empirical risk (as we will see soon, iid structure is very handy). The second term is less than or equal to 0 since $\hat{f}$ minimizes the empirical risk. The last term is an iid sum since f_* is deterministic (not random). As before, we can consider uniform bounds over the feasible set to solve issues that come from non-iid structure.

$$\begin{aligned} R(\hat{f}) - R(f_*) &\le |\hat{R}(\hat{f}) - R(\hat{f})| + 0 + |\hat{R}(f_*) - R(f_*)| \\ &\le \sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)| + 0 + \sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)| \\ \underbrace{R(\hat{f}) - R(f_*)}_{\text{excess risk}} &\le \underbrace{2 \cdot \sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)|}_{\text{empirical process}} \end{aligned} \tag{4.2}$$

The right hand side is called empirical process in statistics. If we can control the empirical process, we can control the generalization error. Intuitively we have bounded the risk of the empirical estimator by the "worst-case" function possible from the function class. We see that the bound in (4.2) translates immediately to

$$\mathbb{P}\big(R(\hat{f}) - R(f_*) \ge \epsilon\big) \le \mathbb{P}\left(\sup_{f\in\mathcal{F}} \Big| \hat{R}(f) - R(f) \Big| \ge \frac{\epsilon}{2} \right), \tag{4.3}$$

where the inequality in (4.3) is due to the fact that if the event $R(\hat{f}) - R(f_*) \ge \epsilon$ happens, since we have (4.2), the event $\sup_{f\in\mathcal{F}} \Big| \hat{R}(f) - R(f) \Big| \ge \frac{\epsilon}{2}$ will also happen.

Uniform convergence generally refers to that the empirical process $\sup_{f\in\mathcal{F}} \Big| \hat{R}(f) - R(f) \Big|$ converges to 0 in probability. Because of (4.3), we see that uniform convergence implies generalization. But we can also talk about explicit convergence rates.

4.2 Finite function classes, $|\mathcal{F}| < \infty$

Our first result in this direction is for finite function classes. Denoting the number of elements in a set with $|\cdot|$, in the following, we assume $|\mathcal{F}| < \infty$.

Theorem 18 (Generalization of Finite Function Classes)**.** *If the function class is finite (i.e. $|\mathcal{F}| < \infty$) and loss is bounded $\ell \le B$, then we have,*

$$\mathbb{P}\left(R(\hat{f}) - R(f_*) < B\sqrt{\frac{2\log(2|\mathcal{F}|) + 2\log\delta^{-1}}{n}} \right) > 1 - \delta. \tag{4.4}$$

Remark.

- The above theorem reads, with probability at least $1-\delta$, we have

  $$R(\hat{f}) - R(f_*) < B\sqrt{\frac{2\log(2|\mathcal{F}|) + 2\log\delta^{-1}}{n}},$$

  This is true in a non-asymptotic sense.

- The complexity measure of the function class $\mathcal{F}$ turns out to be very intuitive in this case, simply the number of functions. The generalization error depends on this quantity in a logarithmic way. This is a good dependence since log grows very slow.

- δ is the confidence level for the bad event. Smaller it is, more risk averse the bound is. It should be chosen in a way that the convergence rate is not affected. In the above bound we observe that $\delta^{-1} = 2|\mathcal{F}|$ is a good choice. The resulting convergence rate is $\mathcal{O}\left(\sqrt{\frac{\log(|\mathcal{F}|)}{n}}\right)$.

- We see that the dependence on number of sample dropped to $\sqrt{n}$ from n (compared to MLE). This is the price we paid to make this result very general, i.e. non-asymptotic and unknown distribution.

- Clearly, this setup doesn't cover any interesting function class since $|\mathcal{F}| < \infty$ almost never holds. For example, think of class of linear functions. How many functions are there in that set?

- The assumption on the loss is also restrictive. It doesn't cover, for example, square loss; yet it does cover 0-1 loss.

Proof. The proof will follow from three steps. In the first step we use a concentration of measure argument for iid averages. In the second, we use the uniform convergence argument derived in (4.3). In the last step, we control the empirical process to obtain a bound on the generalization error. We start with a classical concentration result that will be handy.

Lemma 19 (Hoeffding's inequality)**.** *Suppose $z_1, z_2, \ldots, z_n$ are independent random variables (not necessarily iid) where $a_i \leq X_i \leq b_i$ almost surely. For the partial sums $S_n = n^{-1}\sum_i z_i$ and $\forall \epsilon > 0$, we have*

$$\mathbb{P}\big(|S_n - \mathbb{E}[S_n]| > \epsilon\big) \leq 2 \cdot \exp\left\{ - \frac{2n^2\epsilon^2}{\sum_i (b_i - a_i)^2} \right\}.$$

Remark. The one sided version is also holds without the factor 2 on the right hand side.

1. **Concentration**: We notice that for a non-random f (this excludes $\hat{f}$), $\hat{R}(f) - R(f)$ is the same as $S_n - \mathbb{E}[S_n]$ if we let $z_i \coloneqq \ell((y_i, x_i), f)$. Since loss is bounded by B, by the Hoeffding's inequality, the sample average is concentrating around the true average. That is,

$$\begin{aligned}\mathbb{P}\big(|\hat{R}(f) - R(f)| \geq \epsilon/2\big) &\leq 2 \cdot \exp\left\{ - \frac{n^2\epsilon^2}{2\sum_i B^2} \right\} \\ &\leq 2 \cdot \exp\left\{ - \frac{n\epsilon^2}{2B^2} \right\}\end{aligned}$$

2. **Union bound**: Next, we make use of the finite function class assumption to handle the empirical process.

$$\begin{aligned}\mathbb{P}\left(\sup_{f \in \mathcal{F}} |\hat{R}(f) - R(f)| \geq \epsilon/2 \right) &= \mathbb{P}\left(\bigcup_{f \in \mathcal{F}} \Big\{ |\hat{R}(f) - R(f)| \geq \epsilon/2 \Big\} \right) \\ &\leq \sum_{f \in \mathcal{F}} \mathbb{P}\big(|\hat{R}(f) - R(f)| \geq \epsilon/2\big) \quad \text{(by the union bound)} \\ &\leq 2|\mathcal{F}| \exp\Big\{ - \frac{n\epsilon^2}{2B^2} \Big\}.\end{aligned}$$

3. **Uniform convergence** $\implies$ **Generalization**: Finally, we use the inequality derived in (4.3) to conclude

$$\begin{aligned}\mathbb{P}\big(R(\hat{f}) - R(f_*) \geq \epsilon\big) &\leq \mathbb{P}\left(\sup_{f\in\mathcal{F}} |\hat{R}(f) - R(f)| \geq \epsilon/2\right) \\ &\leq 2|\mathcal{F}| \exp\Big\{-\frac{n\epsilon^2}{2B^2}\Big\} := \delta.\end{aligned}$$

Solving for the δ in the above equation, we obtain

$$\epsilon^2 = \frac{2B^2}{n}\log(2|\mathcal{F}|\delta^{-1}).$$

By substituting $\epsilon(\delta)$ we recover (4.4).

□