1 Motivation

Generalization is a very important problem in machine learning. By saying generalization, we are actually asking how many samples from an unknown distribution $D$ are sufficient to understand the behavior of $D$. The behavior of a distribution $D$ includes many things such as what is the statistics w.r.t class of functions. To answer the question, in this lecture, we introduce Rademacher complexity and its application in a general setting and the supervised learning setting.

2 General Setting

We give the mathematical form of a general setting to analyze generalization. Let $Z$ be the set of all samples and $D$ be the distribution over the set $Z$, i.e. $D \in \mathcal{P}Z$. Assume we do not explicitly know the distribution, but we have the sampler which can generate i.i.d. samples from $D$. $F$ is a function family such that all of its members $f$ are a function from $Z$ to a bounded real value, i.e. $F = \{f : Z \to [-a, a]\}$ where $a$ is a bound on the absolute value of the functions. Note the $|F|$ could be infinite, even uncountably infinite. Our goal is to draw enough samples $S = \{z_1, \cdots, z_m\} \overset{iid}{\sim} D$ that provides good estimations of $E_{z \sim D}[f(z)]$, for all $f \in F$ simultaneously.

We need to emphasize here that estimating all functions $f \in F$ is harder than finding a good estimation for only one fixed $f$. Following are two trivial observations (derived from probability theory). Lemma 1 states how many samples are needed to estimate one statistic. Lemma 2 states the number of samples needed to estimate a finite number of statistics.

**Lemma 1.** For any fixed $f \in F$, if we have data samples $S \triangleq \{z_1, \cdots, z_m\}$ with the size of $m = \Theta\left(\frac{\log(1/\delta)}{\epsilon^2}\right)$, then with probability $\geq 1 - \delta$, the following inequality will hold: $|E_{z \sim S}f(z) - E_{z \sim D}f(z)| \leq \epsilon a$ where $E_{z \sim S}f(x) \triangleq \frac{1}{m} \sum_i f(z_i)$ stands for empirical expectation.

**Lemma 2.** If $|S| \geq \Theta\left(\frac{\log(|F|/\delta)}{\epsilon^2}\right)$, then with probability $\geq 1 - \delta$, $\forall f \in F, |E_{z \sim S}f(z) - E_{z \sim D}f(z)| \leq \epsilon a$.

3 Rademacher Complexity

In the previous section, we raised the question of how many samples are needed to provide good estimates. We also developed answers in the settings where $F$ has finite cardinality. Now we look at how to deal with $F$ with infinite cardinality. Rademacher complexity was developed exactly to answer this question.
**Definition 1** (Empirical Rademacher Complexity (ERC)). Given a sample set $\mathcal{S} = \{z_1, \cdots, z_m\}$, the empirical Rademacher complexity of $\mathcal{F}$ w.r.t $\mathcal{S}$ is defined as

$$\hat{\mathcal{R}}_{\mathcal{S}}(\mathcal{F}) = \mathbb{E}_{\sigma \sim \mathcal{U}\{\pm 1\}^m} \sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i} \sigma_i f(z_i)$$

(1)

where $\mathcal{U}\{\pm 1\}^m$ is a uniform distribution over $m$-dimensional $\{\pm 1\}$ vectors.

Intuitively, ERC measures the expected noise fitting ability of $\mathcal{F}$. As we can see, each term $\frac{1}{m} \sum_{i} \sigma_i f(z_i)$ evaluates the correlation between the noise vector $\sigma$ and the vector $[f(z_1), \cdots, f(z_m)]^T$. By taking the supremum over $\mathcal{F}$, we basically find the best fit to the vector $\sigma$ from our function space.

**Definition 2** (Rademacher Complexity (RC)). The Rademacher complexity of $\mathcal{F}$ w.r.t the distribution $\mathcal{D}$ is defined as,

$$\mathcal{R}_{\mathcal{D},m}(\mathcal{F}) = \mathbb{E}_{\mathcal{S} \sim \mathcal{D}^m}[\hat{\mathcal{R}}_{\mathcal{S}}(\mathcal{F})]$$

(2)

RC is just the average of all ERC over sample set $\mathcal{S}$.

**Theorem 1.** $\forall \delta \in (0, 1) $, $\mathcal{F} \subseteq \{ f : z \rightarrow [a, a+1] \}$ (functions having unit interval value range) and $\mathcal{S} = \{ z_1, \cdots, z_m \}$ i.i.d sampled from $\mathcal{D}$. Then with probability greater than $1 - \delta$,

- $\forall f \in \mathcal{F}, \mathbb{E}_D f(z) \leq \hat{\mathcal{E}}_{\mathcal{S}} f(z) + 2 \mathcal{R}_{\mathcal{D},m}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{m}}$
- $\forall f \in \mathcal{F}, \mathbb{E}_D f(z) \leq \hat{\mathcal{E}}_{\mathcal{S}} f(z) + 2 \hat{\mathcal{R}}_{\mathcal{S}}(\mathcal{F}) + 3 \sqrt{\frac{\log(2/\delta)}{m}}$

Theorem 1 is the main result of how Rademacher complexity bounds the generalization error.

**Remarks 1.** Two-sided bounds hold with slightly weaker constants.

- W.p. $\geq 1 - \delta$, $\forall f \in \mathcal{F}, |\mathbb{E}_D f(z) - \hat{\mathcal{E}}_{\mathcal{S}} f(z)| \leq 2 \mathcal{R}_{\mathcal{D},m}(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{m}}$
- W.p. $\geq 1 - \delta$, $\forall f \in \mathcal{F}, |\mathbb{E}_D f(z) - \hat{\mathcal{E}}_{\mathcal{S}} f(z)| \leq 2 \hat{\mathcal{R}}_{\mathcal{S}}(\mathcal{F}) + 3 \sqrt{\frac{\log(4/\delta)}{m}}$

**Remarks 2.** The bound we get is linear in the value range of the function class $\mathcal{F}$. Suppose $\mathcal{F} \subseteq \{ f : \mathcal{Z} \mapsto [a, b] \}$.

- W.p. $\geq 1 - \delta$, $\forall f \in \mathcal{F}, |\mathbb{E}_D f(z) - \hat{\mathcal{E}}_{\mathcal{S}} f(z)| \leq 2 \mathcal{R}_{\mathcal{D},m}(\mathcal{F}) + (b-a) \sqrt{\frac{\log(2/\delta)}{m}}$
- W.p. $\geq 1 - \delta$, $\forall f \in \mathcal{F}, |\mathbb{E}_D f(z) - \hat{\mathcal{E}}_{\mathcal{S}} f(z)| \leq 2 \hat{\mathcal{R}}_{\mathcal{S}}(\mathcal{F}) + 3(b-a) \sqrt{\frac{\log(4/\delta)}{m}}$

Note that we do not multiply the RC term by the factor $(b-a)$, because the RC itself has linearity. For example, $c\mathcal{R}_{\mathcal{S}}(\mathcal{F}) = \mathcal{R}_{\mathcal{S}}(c\mathcal{F})$ where $c\mathcal{F} \triangleq \{ cf : f \in \mathcal{F} \}$.

**Example.** Let $\mathcal{Z} = \mathbb{R}$, $\mathcal{D}$ is arbitrary, $\mathcal{F} = \{1_{[a,b]}(\cdot) , \forall a \leq b \}$. Suppose $\mathcal{S} = \{ z_1 \leq z_2 \leq \cdots \leq z_m \}$. We compute the ERC $\hat{\mathcal{R}}_{\mathcal{S}}(\mathcal{F}) = \mathbb{E}_\sigma \sup_{a \leq b} \frac{1}{m} \sum_i \sigma_i 1_{[a,b]}(z_i)$). The key observation here is, due to
the discrete property of the $S$, there are only $\Theta(m^2)$ different values of $\sum_i \sigma_{1[a,b]}(z_i)$. For a fixed $(a,b)$ pair, we claim the following.

$$
\mathbb{P}_\sigma \left[ \sum_{a \leq z_i \leq b} \sigma_i \geq c \sqrt{m \log m} \right] \leq \frac{1}{m^3} \Rightarrow \mathbb{P}_\sigma \left[ \exists a,b \left| \sum_{a \leq z_i \leq b} \sigma_i \right. \geq c \sqrt{m \log m} \right] \leq \frac{1}{m}
$$

(3)

Then we can use this probability bound to bound the Rademacher complexity as follows. Let $A$ be the event that $\exists a,b \left| \sum_{a \leq z_i \leq b} \sigma_i \right. \geq c \sqrt{m \log m}$ happens.

$$
\hat{\mathcal{R}}_S(F) = \mathbb{E}_\sigma \left[ \sup_{a \leq b} \frac{1}{m} \sum_i \sigma_{1[a,b]}(z_i) \right] = \mathbb{P}(A) \cdot \mathbb{E}_\sigma \left[ \sup_{a \leq b} \frac{1}{m} \sum_i \sigma_{1[a,b]}(z_i) \mid A \right] + (1 - \mathbb{P}(A)) \cdot \mathbb{E}_\sigma \left[ \sup_{a \leq b} \frac{1}{m} \sum_i \sigma_{1[a,b]}(z_i) \mid A^C \right]
$$

(4)

$$
\leq \frac{1}{m} \sup_{a \leq b, \sigma} \frac{1}{m} \left| \sum_i \sigma_{1[a,b]}(z_i) \right| + 1 \cdot \frac{1}{m} c \sqrt{m \log m} \leq \mathcal{O}\left( \sqrt{\frac{\log m}{m}} \right)
$$

(5)

(6)

With $\hat{\mathcal{R}}_S(F) \leq \mathcal{O}\left( \sqrt{\frac{\log m}{m}} \right)$ combining Theorem 1 we get the following generalization bound, if $S \sim D$, then w.p. $\geq 1 - \delta$, $\forall [a,b] : \left| D([a,b]) - \hat{D}_S([a,b]) \right| \leq \mathcal{O}\left( \sqrt{\frac{\log m + \log 1/\delta}{\sqrt{m}}} \right)$ where $D([a,b]) = \mathbb{E}_{z \sim D} 1_{a \leq z \leq b} = \mathbb{P}_D(a \leq z \leq b)$ is the probability that a random variable with distribution $D$ falls in the interval $[a,b]$ while $\hat{D}_S$ is the empirical distribution defined by the sample set $S$. So actually, the uniform error bound of $\left| D([a,b]) - \hat{D}_S([a,b]) \right|$ indicates that the CDF of the distribution $D$ and the CDF of the empirical distribution $\hat{D}_S$ are very close. In another word, we can learn the CDF of a distribution within $\epsilon$ Kolmogorov distance, from $\mathcal{O}(\frac{1}{\epsilon^2} \log \frac{1}{\epsilon})$ samples.

**Properties of Rademacher complexity.**

1. $F' \subseteq F \Rightarrow \hat{\mathcal{R}}_S(F') \leq \hat{\mathcal{R}}_S(F)$.

2. For a fixed function $h : \mathcal{X} \mapsto \mathbb{R}$, $\hat{\mathcal{R}}_S(F + h) = \hat{\mathcal{R}}_S(F)$.

3. Suppose $cuv$ is a convex hull, $\hat{\mathcal{R}}_S(cuv(F)) = \hat{\mathcal{R}}_S(F)$.

4. $\hat{\mathcal{R}}_S(F' + F) = \hat{\mathcal{R}}_S(F') + \hat{\mathcal{R}}_S(F)$ and $\hat{\mathcal{R}}_S(cF) = c \hat{\mathcal{R}}_S(F)$.

5. Suppose $\phi_1, \cdots, \phi_m : \mathbb{R} \mapsto \mathbb{R}$ are $c$-lipschitz, $\mathbb{E}_\sigma \sup_{f \in F} \sum_i \sigma_i \phi_i(f(z_i)) \leq c \cdot \mathbb{E}_\sigma \sup_{f \in F} \sum_i \sigma_i f(z_i)$.

6. Suppose $l : \mathbb{R} \mapsto \mathbb{R}$ is $c$-lipschitz, $\hat{\mathcal{R}}_S(l \circ F) \leq c \hat{\mathcal{R}}_S(F)$.

7. (Massarts Lemma) $\hat{\mathcal{R}}_S(F) \leq (b - a) \sqrt{\frac{2 \log |F|}{m}}$.

**Proof.** Proof for properties 1,2,3,4 are trivial via using the definition. The proof of Massart Lemma can be found in the book [2](section 3).
4 Supervised Learning Setting

Now let us apply the Rademacher complexity in the supervised learning setting. In supervised learning, we have,

- The data space $X \times Y$ (feature and label) and a data distribution $D$ over it.
- Concept classes (hypotheses) $\mathcal{H} \subseteq \{ h : X \mapsto Y \}$.
- Loss function $l : Y \times Y \mapsto \mathbb{R}$ measuring the error between the predicted label and the ground truth. For example, in the classification task, loss can be $l(y, y') = 1_{[y \neq y']}$. While in the regression task, loss can be $l(y, y') = (y - y')^2$.

Our goal is to find the best model (classifier) that minimize the loss, i.e. $\min_{h \in \mathcal{H}} \mathbb{E}_{(x,y) \sim D}[l(h(x), y)]$.

The question we ask about generalization is how many i.i.d samples $(x_1, y_1), \cdots, (x_m, y_m) \overset{\text{iid}}{\sim} D$ are needed to guarantee that the empirical loss of a model is very close to the inaccessible real loss?

**Reduce to general setting.** In fact, we can reduce the supervised learning setting to the general setting by defining the function class as following, $\mathcal{F} = \{ (x, y) \mapsto l(h(x), y) \mid \forall h \in \mathcal{H} \}$. Directly applying Theorem 1 we will get the generalization bounds for supervised learning.

**Theorem 2.** If $\mathcal{S} = \{ (x_1, y_1), \cdots, (x_m, y_m) \overset{\text{iid}}{\sim} D \}$, then w.p. $\geq 1 - \delta$,

$$\forall f \in \mathcal{F}, \quad |\mathbb{E}_D[l(h(x), y)] - \hat{\mathbb{E}}_S[l(h(x), y)]| \leq 2\hat{R}_S(\mathcal{F}) + 3\sqrt{\frac{\log(4/\delta)}{m}}.$$  

(7)

**An useful observation** is that, if $Y \subseteq \mathbb{R}$ and $l(\cdot, y)$ is $c$–Lipschitz in its first argument for all $y \in Y$, then

$$\hat{R}_S(\mathcal{F}) = \frac{1}{m}\mathbb{E}_\sigma[\sup_{h \in \mathcal{H}} \sum_i \sigma_i l(h(x_i), y_i)] \leq \frac{c}{m}\mathbb{E}_\sigma[\sup_{h \in \mathcal{H}} \sum_i \sigma_i h(x_i)] = c \cdot \hat{R}_S(\mathcal{H}).$$  

(8)

The inequality can be directly got via the fifth properties of Rademacher complexity as we introduced in Section 3.

Now let us see two examples of generalization analysis for machine learning models.

**Example 1.** Lasso (linear regression with $l_1$ regularization)

Setting: given $(x_1, y_1), \cdots, (x_m, y_m) \in \mathbb{R}^d \times \mathbb{R}$, we want $\min_{f} \sum_i (y_i - f^T x_i)^2$, s.t. $\|f\|_1 \leq R$. The coefficients $f \in \mathbb{R}^d$ are learnt to fit the data while the convex constrain $\|f\|_1 \leq R$ promotes the sparsity of $f$.

Generalization Analysis: To answer the question that does the empirical $l_2^2$ loss, defined by the data $(x_1, y_1), \cdots, (x_m, y_m)$ that Lasso optimizes over, really match the real loss under $D$, we need to bound the Rademacher complexity of the function class $\mathcal{F}$.

$$\mathcal{F} = \{ (x, y) \mapsto (y - f^T x)^2, \text{ where } \|f\|_1 \leq R \}.$$  

(9)

**Lemma 3.** Assuming that $D$ only outputs $(x, y)$ in a bounded set $[-B, B]^d$ then,

$$\hat{R}_S(\mathcal{F}) \leq \mathcal{O}\left( R^2 B^2 \sqrt{\frac{\log(2d)}{m}} \right).$$
Proof. The key observation here is, for bounded data \((x, y) \in [-B, B]^{d+1}\), the squared \(l^2\) loss function \((y_i - \cdot)^2\) will be \(2(R + 1)B\)-Lipschitz for every fixed \(y_i\). Let \(\mathcal{G} = \{ x \mapsto f^T x : f \in \mathbb{R}^d, \|f\|_1 \leq R \}\).

\[
\hat{\mathcal{R}}_S(\mathcal{G}) = \frac{1}{m} \mathbb{E}_\sigma[\sup_{f : \|f\|_1 \leq R} f^T \sum_i \sigma_i x_i] = \frac{R}{m} \mathbb{E}_\sigma[\sup_{f : \|f\|_1 \leq 1} f^T \sum_i \sigma_i x_i]
\]

\[
= R \cdot \hat{\mathcal{R}}_S(cvs(\{e_1, -e_1, \cdots, e_d, -e_d\})) = R \cdot \hat{\mathcal{R}}_S(\{e_1, -e_1, \cdots, e_d, -e_d\})
\]

\[
\leq R \cdot 2B \sqrt{\frac{2\log(2d)}{m}}
\]

\[
\Rightarrow \hat{\mathcal{R}}_S(\mathcal{F}) \leq 2(R + 1)B \cdot \hat{\mathcal{R}}_S(\mathcal{G}) \leq O\left(R^2B^2\sqrt{\frac{\log(2d)}{m}}\right)
\]

The inequality connection between \(\hat{\mathcal{R}}_S(\mathcal{F})\) and \(\hat{\mathcal{R}}_S(\mathcal{G})\) comes from the composition property (property 5) of the Rademacher complexity as we emphasized in Eqn. 8. While in the derivation of the sample complexity \(\hat{\mathcal{R}}_S(\mathcal{G})\), the last inequality comes from Massart’s Lemma (property 7 of the Rademacher complexity) and the last equality comes from property 3 of the Rademacher complexity.

The result stated in Lemma 3 is pretty cool since the Rademacher complexity is logarithmic in \(d\) which means that we can still have a good generalization bound even when the data dimension is very high.

**Example 2. Neural Networks (Multilayer Perceptron)**

We denote \(\mathcal{F}_k\) to the function class of a \(k\) layer neural network. Then \(\mathcal{F}_k\) can be recursively defined as follows

\[
\mathcal{F}_k = \{ x \mapsto \sum_j w_j^k \eta(f_j(x)) \mid \forall j, f_j \in \mathcal{F}_{k-1}, \|w^k\|_1 \leq B_k \}
\]

where \(\eta\) is an activation function such as a sigmoid or ReLU and \(w^k\) are the neural network weights in the \(k^{th}\) layer.

**Lemma 4.** Assume the activation function \(\eta\) is 1-Lipchitz and weights are bounded, i.e. \(\|w^i\|_1 \leq B_i\). Then \(\hat{\mathcal{R}}_S(\mathcal{F}_k) \leq (\prod_i 2B_i)\hat{\mathcal{R}}_S(\mathcal{F}_1)\).
Proof.

\[
\hat{R}_S(F_k) = \frac{1}{m} \mathbb{E}_\sigma \left[ \sup_{j, f_j \in \mathcal{F}_{k-1}, \|w^k\|_1 \leq B_k} \sum_{i=1}^m \sigma_i w^k_j \eta(f_j(x_i)) \right] 
\]

(15)

\[
\leq \frac{1}{m} \mathbb{E}_\sigma \left[ \sup_{j, f_j \in \mathcal{F}_{k-1}, \|w^k\|_1 \leq B_k} \|w^k\|_1 \max_j \left| \sum_i \sigma_i \eta(f_j(x_i)) \right| \right] 
\]

(16)

\[
\leq \frac{B_k}{m} \mathbb{E}_\sigma \left[ \sup_{j, f_j \in \mathcal{F}_{k-1}} \left| \sum_i \sigma_i \eta(f_j(x_i)) \right| \right] 
\]

(17)

\[
\leq \frac{2B_k}{m} \mathbb{E}_\sigma \left[ \sup_{f \in \mathcal{F}_{k-1}} \left| \sum_i \sigma_i \eta(f(x_i)) \right| \right] 
\]

(18)

\[
= 2B_k \hat{R}_S() \leq 2B_k \hat{R}_S(F_{k-1}) 
\]

(19)

The second inequality holds because of bounded weights. The last inequality holds because of the Lipchitz continuity assumption of the activation function.

\[
= 2B_k \hat{R}_S(F_{k-1}) \leq 2B_k \hat{R}_S(F_{k-1}) 
\]

(20)

5 Discussion

In previous sections, we motivated the generalization problem and introduced the generalization analysis based on the Rademacher complexity. The generalization theory works very well to analyze simple machine learning models such as SVMs. However, in the new era of deep learning, more challenges emerge. In fact, deep learning has been characterized by having a large number of parameters (often over-parameterized) and yet has surprisingly well generalization performance. The generalization error of neural networks is much smaller than that suggested by its huge Rademacher complexity. Thus there is more to be discovered to understand the generalization behavior of deep neural networks. In the machine learning community, researchers have started rethinking about generalization for deep learning [3].

As discussed in the class, one possible way to explain the good generalization behavior of NNs could be the inductive bias of SGD, the most common way to optimize NNs. Some recent work [1] shows that by using SGD, a large over-parameterized NN actually will converge to a model that matches the behavior of a tiny size NN. In other words, the model family that is reachable after SGD converges might be extremely smaller than the model family defined by its parameters. Thus the actual Rademacher complexity of the NNs trained by SGD could be much smaller than we thought. At a high level, the neural network has two strategies to fit the training data, ‘memorizing’ (which does not generalize) or ‘learning something’ (which generalizes). The optimization process would be very critical to decide what strategy the neural network uses.

References

[1] ALLEN-ZHU, Z., LI, Y., AND LIANG, Y. Learning and generalization in overparameterized
