**CS 7545: Machine Learning Theory** **Fall 2019**

# Lecture 19: Statisical Learning Theory

*Lecturer: Jacob Abernethy* *Scribes: Disha Das, Gregory Hessler*



## 19.1 Supervised Learning

The following are key ingredients for a supervised statistical learning scenario

1. An observation space $\mathcal{X}$
2. A label space $\mathcal{Y}$
   Examples:
   - $\{0, 1\}$ "classification"
   - $[k]$ "multi-class clasification"
   - $\mathbb{R}$ "regression"
3. A prediction space $\hat{\mathcal{Y}}$. Often this is the same as the label space.
   Example where they differ:
   - $\mathcal{Y} = \{0, 1\}$ and $\hat{\mathcal{Y}} = [0, 1]$
4. An unknown distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$
5. A hypothesis space $\mathcal{H}$
   Examples:
   - Linear threshold functions: $\mathcal{H} = \{h_{\mathbf{w},b}(\mathbf{x}) = \mathbb{1}[\mathbf{w} \cdot \mathbf{x} + b > 0]\}$
   - Decision stumps: $\mathcal{H} = \{h_{i,c}(\mathbf{x}) = \mathbb{1}[\mathbf{x}_i > c]\}$
   - Neural Networks: $\mathcal{H} = \{h_{M_1,b_1,M_2,b_2,\ldots M_k,b_k}(\mathbf{x}) = \sigma(b_k + M_k\sigma(b_{k-1} + M_{k-1}\sigma(\ldots(\mathbf{x})))\}$ where $\sigma$ is the sigmoid function
6. A loss function: $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$
   Examples
   - $\ell(\hat{y}, y) = (y - \hat{y})^2$ "squared loss" (often used for regression)
   - $\ell(\hat{y}, y) = \max(0, 1 - \hat{y}y)$ "hinge loss"
   - $\ell(\hat{y}, y) = \mathbb{1}[\hat{y} \neq y]$ "0-1 loss"

### 19.1.1 Risk and Empirical Risk

**Definition 19.1 (Risk)** *Given a distribution $\mathcal{D}$, a hypothesis $h \in \mathcal{H}$, and a loss function $\ell$, we define the* ***risk*** *of $h$ as*

$$\mathcal{R}(h) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(h(\mathbf{x}), y)].$$

Note that we typically cannot compute $\mathcal{R}(\cdot)$ as we would need an infinite amount of data.

**Definition 19.2 (Empirical Risk)** *Given data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\} \sim \mathcal{D}$, the* ***empirical risk*** *of a hypothesis $h$ is defined as*

$$\hat{\mathcal{R}}_n(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i, y_i).$$

## 19.2 Empirical Risk Minimization

**Definition 19.3 (Empirical Risk Minimization (ERM))** *The* ***Empirical Risk Minimization*** *algorithm "learns" a best hypothesis* $\hat{h}_n^{ERM}$ *which minimizes the empirical risk*

$$\hat{h}_n^{ERM} = \arg\min_{h\in\mathcal{H}} \hat{\mathcal{R}}_n(h).$$

Once we have found $\hat{h}_n^{\text{ERM}}$, we often wish to know how well it "generalizes" through metrics such as the estimation error and approximation error

**Definition 19.4 (Estimation Error)** *We define the* ***estimation error*** *of an ERM hypothesis* $\hat{h}_n^{ERM}$ *as*

$$\mathcal{R}(\hat{h}_n^{ERM}) - \min_{h^*\in\mathcal{H}} \mathcal{R}(h^*).$$

**Definition 19.5 (Approximation Error)** *We define the* ***approximation error*** *of an ERM hypothesis* $\hat{h}_n^{ERM}$ *as*

$$\mathcal{R}(h^*) - \min_{\textit{all functions } h^{**}} \mathcal{R}(h^{**}).$$

### 19.2.1 Bounding the Estimation Error

Notice that

$$\mathcal{R}(\hat{h}_n^{ERM}) - \min_{h\epsilon\mathcal{H}} \mathcal{R}(h^*) = \tag{19.1}$$

$$\mathcal{R}(\hat{h}_n^{ERM}) - \hat{\mathcal{R}}_n(\hat{h}_n^{ERM}) \tag{T$_1$}$$

$$+\hat{\mathcal{R}}_n(\hat{h}_n^{ERM}) - \hat{\mathcal{R}}_n(h^*) \tag{T$_2$}$$

$$+\hat{\mathcal{R}}_n(h^*) - \mathcal{R}(h^*) \tag{T$_3$}$$

**Claim 19.6** $T_2 \leq 0$

**Proof:** The above follows from the definition of $\hat{h}_n^{ERM}$

$$\hat{h}_n^{\text{ERM}} = \arg\min_{h\in\mathcal{H}} \hat{\mathcal{R}}_n(h).$$

∎

**Claim 19.7** $T_1$ *and* $T_3 \leq \sup_{h\in\mathcal{H}} |\mathcal{R}(h) - \hat{\mathcal{R}_n}(h)|$

**Remark:** Bounding the above quantity $\sup_{h\in\mathcal{H}} |\mathcal{R}(h) - \hat{\mathcal{R}_n}(h)|$ is known as a **Uniform Deviation Bound**. Here, $\hat{\mathcal{R}_n}(h)$ corresponds to the *training error* and $\mathcal{R}(h)$ corresponds to the *test error*.

**Proof:** Inorder to bound $\sup_{h\in\mathcal{H}} |\mathcal{R}(h) - \hat{\mathcal{R}_n}(h)|$, we try to prove by incorrect derivation. Let

$$\mathcal{Z}_i = l(\hat{h}_n^{ERM}(x_i), y_i)$$

where $(x_i, y_i)$ is the $i^{th}$ sample of training set.

$$\mathcal{R}(\hat{h}_n^{ERM}) = \mathbb{E}_{(x,y)\sim D}[l(\hat{h}_n^{ERM}(x), y)] = \mu$$

So,

$$\mathbb{E}_{(x_i,y_i)}[\mathcal{Z}_i] = \mu$$

Using Hoeffding's inequality (assume $l$ is bounded in $[0, 1]$)

$$\hat{\mathcal{R}}_n(\hat{h}_n^{ERM}) - \mathcal{R}(\hat{h}^{ERM}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{Z}_i - \mu \leq \sqrt{\frac{log 1/\delta}{2n}}$$

with probability $\geq 1 - \delta$. But this is incorrect. If we get rid of "ERM", everything will be fine. But with ERM, this claim is not true. This is because when we bound with Hoeffding's inequality, we require $\mathcal{Z}_i$ to be independent. However, the ERM hypothesis $\hat{h}_n^{ERM}$ makes the samples correlated and violates our assumption. Hence the above derivation is incorrect.

The following can be done:

$$\begin{aligned}
&Pr(|\hat{\mathcal{R}}_n(\hat{h}) - \mathcal{R}(\hat{h})| > t) \\
&\leq Pr(\exists h \in \mathcal{H} : |\hat{\mathcal{R}}_n(\hat{h}) - \mathcal{R}(h)| > t) \\
&\leq \sum_{h \in \mathcal{H}} Pr(|\hat{\mathcal{R}}_n(\hat{h}) - \mathcal{R}(h)| > t) \\
&\leq |\mathcal{H}| exp(-2nt^2) = \delta
\end{aligned}$$

Thus,

$$Pr(|\hat{\mathcal{R}}_n(\hat{h}) - \mathcal{R}(\hat{h})| > t) \leq |\mathcal{H}| exp(-2nt^2)$$

With probability atleast $1 - \delta$

$$|\hat{R}_n(\hat{h}_n) - \mathcal{R}(\hat{h})| \leq \sqrt{\frac{log|\mathcal{H}|/\delta}{2n}}$$

■