CMSC 35900 (Spring 2008) Learning Theory **Lecture: 8**

Concentration, ERM, and Compression Bounds

Instructors: Sham Kakade and Ambuj Tewari

1 Chernoff and Hoeffding Bounds

Theorem 1.1. *Let $Z_1, Z_2, \ldots Z_m$ be m i.i.d. random variables with $Z_i \in [a, b]$ (with probability one). Then for all $\epsilon > 0$ we have:*

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} Z_i - \mathbb{E}[Z] > \epsilon\right) \leq e^{-\frac{2m\epsilon^2}{(b-a)^2}}$$

The union bound states that for events $C_1, C_2, \cdots C_m$ we have:

$$\mathbb{P}(C_1 \cup C_2 \ldots \cup C_m) \leq \sum_{i=1}^{m} \mathbb{P}(C_i)$$

which holds for all events. If the events are C_i exclusive, then we have equality:

$$\mathbb{P}(C_1 \cup C_2 \ldots \cup C_m) = \sum_{i=1}^{m} \mathbb{P}(C_i)$$

Typically, the union bound introduces much slop into our bounds (though it is used often as understanding dependencies is often tricky).

2 Empirical Risk Minimization (ERM)

Suppose we have a training data set $(X_1, Y_1), \ldots, (X_m, Y_m)$ consisting of independent and identically distributed random variable pairs from an unknown probability distribution.

For any hypothesis $f \in \mathcal{F}$, we know that $\phi(f(X_i), Y_i)$ is an unbiased estimate of the risk $L_\phi(f)$. Hence, we know that:

$$\hat{L}_\phi(f) = \frac{1}{m}\sum_{i=1}^{m} \phi(f(X_i), Y_i)$$

is also an unbiased estimate of $L_\phi(f)$.

The ERM algorithm is to choose the hypothesis which minimizes this empirical risk, i.e.

$$\hat{f} = \text{argmin}_{f \in \mathcal{F}} \frac{1}{m}\sum_{i=1}^{m} \phi(f(X_i), Y_i)$$

Two central questions are in bounding

$$|L_\phi(f) - \hat{L}_\phi(\hat{f})| \leq ??$$

and

$$L_\phi(\hat{f}) - L_\phi(f^*) \leq ??$$

The former is how much our estimate differs from the best. The latter is how close the risk of our hypothesis is to that of the optimal hypothesis.

3 Generalization Bounds for the Finite Case

Now let us consider the case where $\mathcal{F}$ is finite and the loss is bounded in $[0,1]$

Here we have that:

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\hat{L}_\phi(f)-L_\phi(f)\right|\geq\epsilon\right) &= \mathbb{P}\left(\exists f\in\mathcal{F}\text{ s.t. }|L(f)-\hat{L}(f)|\geq\epsilon\right)\\ &\leq \sum_{f\in\mathcal{F}}\mathbb{P}\left(|L(f)-\hat{L}(f)|\geq\epsilon\right)\\ &\leq 2|\mathcal{F}|e^{-2m\epsilon^2}\end{aligned}$$

Now if we apriori choose

$$\epsilon=\sqrt{\frac{\log 2|\mathcal{F}|+\log\frac{1}{\delta}}{2m}}$$

then we have

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\left|\hat{L}_\phi(f)-L_\phi(f)\right|\geq\sqrt{\frac{\log 2|\mathcal{F}|+\log\frac{1}{\delta}}{2m}}\right)\leq\delta$$

Equivalently, this says that with probability greater than $1-\delta$, for all $f\in\mathcal{F}$

$$\left|\hat{L}_\phi(f)-L_\phi(f)\right|\leq\sqrt{\frac{\log 2|\mathcal{F}|+\log\frac{1}{\delta}}{2m}}$$

which is a *uniform convergence* statement. And this implies the following performance bound of ERM:

$$L_\phi(\hat{f})\leq L_\phi(f^*)+2\sqrt{\frac{\log 2|\mathcal{F}|+\log\frac{1}{\delta}}{2m}}$$

Note the logarithmic dependence on the size of the hypothesis class.

4 Occam's Razor Bound

Now consider partitioning the error probability δ to each $f\in\mathcal{F}$. In particular, assume we have specified a δ_f for each $f\in\mathcal{F}$ such that:

$$\sum_{f\in\mathcal{F}}\delta_f\leq\delta$$

The following theorem is referred to as the "Occam's Razor Bound"

Theorem 4.1. *Equivalently, this says that with probability greater than $1-\delta$, for all $f\in\mathcal{F}$*

$$\left|\hat{L}_\phi(f)-L_\phi(f)\right|\leq\sqrt{\frac{\log\frac{2}{\delta_f}}{2m}}$$

which is a uniform convergence *statement.*

Proof. Define:

$$\epsilon_f=\sqrt{\frac{\log\frac{2}{\delta_f}}{2m}}$$

We have that:

$$\begin{aligned}
\mathbb{P}\left(\exists f \in \mathcal{F} \text{ s.t. } |L(f) - \hat{L}(f)| \geq \epsilon_f\right) &\leq \sum_{f \in \mathcal{F}} \mathbb{P}\left(|L(f) - \hat{L}(f)| \geq \epsilon_f\right) \\
&\leq \sum_{f \in \mathcal{F}} 2e^{-2m\epsilon_f^2} \\
&= \sum_{f \in \mathcal{F}} \delta_f \\
&\leq \delta
\end{aligned}$$

which completes the proof. □

5 Compression Bound for the Realizable Case

Now let us consider a different type of algorithm, where we do not apriori explicitly define the hypothesis class. Here, let T be ordered training set — we consider the training set as the *ordered sequence*:

$$(X_1, Y_1), \ldots, (X_m, Y_m) \ .$$

The learning algorithm $\mathcal{A}$ takes as input T and returns a hypothesis f.

Now let us consider a special case where our algorithm would provide the same output as $\mathcal{A}(T)$ if it had been given as input only a subsequence of T. More precisely, let $I \subset [m]$. For the increasing subsequence $i_1\, i_2, \ldots i_l$, where $i_j \in I$ and $l = |I|$ (this just lists all of I in increasing order), define the corresponding subsequence of T as:

$$T_I = (X_{i_1}, Y_{i_1}), (X_{i_2}, Y_{i_2}) \ldots, (X_{i_l}, Y_{i_l}) \ .$$

So T_{-I} denotes the subsequence corresponding to $-I$ (the complement of I in $[m]$). Now we say that I is a compression set for T if:

$$\mathcal{A}(T) = \mathcal{A}(T_I)$$

Intuitively, if I is small and the empirical risk of $\mathcal{A}(T)$ is small, then we would expect that our hypothesis has good performance.

For example, let us run the perceptron algorithm on T and let $\mathcal{A}(T)$ be the final weight vector returned after the algorithm is run on T. Here, a compression set is:

$$I = \{ \text{ the times t at which the perceptron algorithm made a mistake } \}$$

By definition of the perceptron algorithm, we know that $\mathcal{A}(T)$ is equal to $\mathcal{A}(T_I)$ so I indeed is a compression set.

For the following theorem, it is useful to define the empirical error on an index set I as:

$$\hat{L}_I(f) = \frac{1}{|I|} \sum_{i \in I} \phi(f(X_i), Y_i)$$

Now we are ready to state the compression bound.

Theorem 5.1. *(Compression Bound Realizable Case) Assuming that the loss is bounded in* $[0, 1]$. *With probability at least* $1 - \delta$*, we have that if I is a compression set for* T*, and* $\hat{L}_{-I}(\mathcal{A}(T)) = 0$*, then:*

$$L(\mathcal{A}(T)) \leq \frac{1}{m - l}\left((l + 1)\log m + \log \frac{1}{\delta}\right)$$

where l *is the size of the compression set and the probability is with respect to a random draw of* T*.*

Proof. The event we seek to bound the probability of is:

$$\exists I \text{ s.t. I is a compression set for T} \wedge \hat{L}_{-I}(\mathcal{A}(T_I)) = 0 \wedge L(\mathcal{A}(T_I)) \geq \epsilon$$

for an appropriate choice of ϵ.

We start by bounding the probability that this event occurs for some fixed compression set size l. We will take a union bound over l later.

$$\begin{aligned}
& \mathbb{P}\left(\exists I \text{ s.t. } |I| = l \wedge \text{I is a compression set for T} \wedge \hat{L}_{-I}(\mathcal{A}(T_I)) = 0 \wedge L(\mathcal{A}(T_I)) \geq \epsilon\right) \\
\leq \;& \mathbb{P}\left(\exists I \text{ s.t. } |I| = l \wedge \hat{L}_{-I}(\mathcal{A}(T_I)) = 0 \wedge L(\mathcal{A}(T_I)) \geq \epsilon\right) \\
\leq \;& \sum_{I \subset [m] \text{ s.t. } |I| = l} \mathbb{P}\left(\hat{L}_{-I}(\mathcal{A}(T_I)) = 0 \wedge L(\mathcal{A}(T_I)) \geq \epsilon\right) \\
= \;& \sum_{I \subset [m] \text{ s.t. } |I| = l} \mathbb{E}\left[\mathbb{P}_{T_{-I}}\{\hat{L}_{-I}(\mathcal{A}(T_I)) = 0 \wedge L(\mathcal{A}(T_I)) \geq \epsilon\} \Big| T_I\right]
\end{aligned}$$

Now for any fixed T_I, the last prob is just the probability of having a true risk greater than ϵ and an empirical risk of 0 on a test set of size $m - l$.

Now for any random variable $z \in [0, 1]$ (with probability one), if $\mathbb{E}[z] \geq \epsilon$ then $\mathbb{P}(z = 0) \leq 1 - \epsilon$. Hence, for a given T_I we have that:

$$\mathbb{P}_{T_{-I}}\{\hat{L}_{-I}(\mathcal{A}(T_I)) = 0 \wedge L(\mathcal{A}(T_I)) \geq \epsilon\} \leq (1 - \epsilon)^{m-l}$$

by the binomial tail bound. Proceeding we have:

$$\begin{aligned}
\leq \;& \sum_{I \subset [m] \text{ s.t. } |I| = l} (1 - \epsilon)^{m-l} \\
\leq \;& m^l (1 - \epsilon)^{m-l} \\
\leq \;& m^l e^{-(m-l)\epsilon} \quad .
\end{aligned}$$

If we desire that this probability is less than δ/m then an appropriate setting of ϵ is:

$$\epsilon = \frac{1}{m - l}\left((l + 1)\log m + \log\frac{1}{\delta}\right) \quad .$$

which can be seen by solving for ϵ in the above equation.

To complete the proof:

$$\begin{aligned}
& \mathbb{P}\left(\exists I \text{ s.t. I is a compression set for T} \wedge \hat{L}_{-I}(\mathcal{A}(T_I)) = 0 \wedge L(\mathcal{A}(T_I)) \geq \epsilon\right) \\
\leq \;& \sum_l \mathbb{P}\left(\exists I \text{ s.t. } |I| = l \wedge \text{I is a compression set for T} \wedge \hat{L}_{-I}(\mathcal{A}(T_I)) = 0 \wedge L(\mathcal{A}(T_I)) \geq \epsilon\right) \\
\leq \;& \sum_l \delta/m \\
= \;& \delta
\end{aligned}$$

where we have used the union bound. □