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 $\varepsilon > 0$ we have:

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

The union bound states that for events $C_1, C_2, \ldots C_m$ 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 dependen-

cies 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.

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

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 \( F \) is finite and the loss is bounded in \([0, 1]\).

Here we have that:
\[
P \left( \sup_{f \in F} \left| \hat{L}_\phi(f) - L_\phi(f) \right| \geq \varepsilon \right) = P \left( \exists f \in F \text{ s.t. } |L(f) - \hat{L}(f)| \geq \varepsilon \right) \leq \sum_{f \in F} P \left( |L(f) - \hat{L}(f)| \geq \varepsilon \right) \leq 2|F|e^{-2m\varepsilon^2}
\]

Now if we apriori choose
\[
\varepsilon = \sqrt{\frac{\log 2|F| + \log \frac{1}{\delta}}{2m}}
\]
then we have
\[
P \left( \sup_{f \in F} \left| \hat{L}_\phi(f) - L_\phi(f) \right| \geq \sqrt{\frac{\log 2|F| + \log \frac{1}{\delta}}{2m}} \right) \leq \delta
\]

Equivalently, this says that with probability greater than \( 1 - \delta \), for all \( f \in F \)
\[
\left| \hat{L}_\phi(f) - L_\phi(f) \right| \leq \sqrt{\frac{\log 2|F| + \log \frac{1}{\delta}}{2m}}
\]
which is a uniform convergence statement. And this implies the following performance bound of ERM:
\[
L_\phi(\hat{f}) \leq L_\phi(f^*) + 2\sqrt{\frac{\log 2|F| + \log \frac{1}{\delta}}{2m}}
\]

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

4 Occam’s Razor Bound

Now consider partitioning the error probability \( \delta \) to each \( f \in F \). In particular, assume we have specified a \( \delta_f \) for each \( f \in F \) such that:
\[
\sum_{f \in F} \delta_f \leq \delta
\]

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 $f \in \mathcal{F}$

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

which is a uniform convergence statement.

Proof. Define:

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

We have that:

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

which completes the proof. \qed