Chapter 1

Learning via uniform convergence

In this Chapter we will develop a more general tool, namely the *uniform convergence*, and apply it to show that any finite class is learnable in the agnostic PAC model with general loss functions, as long as the range loss function is bounded.

1.1 Uniform convergence is sufficient for learnability

Given a hypothesis class $\mathcal{H}$, the ERM learning algorithm:

- Receives a training sample S.
- The learner evaluates the risk of each $h \in \mathcal{H}$ on the given sample S.
- The learner outputs a member h^* of $\mathcal{H}$ that minimizes this empirical risk.
- The hope is that a h that minimizes the empirical risk with respect to S is a risk minimizer with respect to the true data probability distribution as well. For that, it sufficies to ensure that the empirical risks of all members of $\mathcal{H}$ are good approximations of their true risk.

In other words, we need that uniformly over all hypoteses in the hypotesis class, the empirical risk will be close to the true risk. This is formalized in the following definition. Then, a lemma introduces us to a first result.

Definition 1.1.1 (ε-representative sample)**.** A training set S is called ε-representative (with respect to domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if:

$$\forall h \in \mathcal{H},\ |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon \tag{1.1}$$

Lemma 1.1.1. *Assume that a training set S is $\frac{\varepsilon}{2}$-representative (with respect to domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$). Then any output of ERM$_{\mathcal{H}}(S)$, namely, any $h_S \in \operatorname{argmin}_{h\in\mathcal{H}} L_S(h)$, satisfies:*

$$L_{\mathcal{D}}(h_S) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon \tag{1.2}$$

Proof. For every $h \in \mathcal{H}$:

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\varepsilon}{2} \leq L_S(h) + \frac{\varepsilon}{2} \leq L_{\mathcal{D}}(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = L_{\mathcal{D}}(h) + \varepsilon$$

□

The consequence of this lemma is that, if with probability $1 - \delta$, a random training set S is ε-representative, then the ERM rule is an agnostic PAC learner. The uniform convergence condition formalizes this requirement.

Definition 1.1.2 (Uniform convergence)**.** A hypothesis calss $\mathcal{H}$ has the *uniform convergence property* (with respect to a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{\mathrm{UC}} : (0,1)^2 \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1 - \delta$, S is ε-representative.