1 Sample Size Analysis

In this section we examine the required sample sizes for generalization from training domains, with finite sample sizes, to new unseen domains. We start with some notations:

- We recall the definition of performance indicator:

$$\mathbb{1}_{\mathcal{L},\gamma}(h,e) = \mathbb{1}[\mathcal{L}(h,e) \geq \gamma]$$

- We denote the training set, that is gathered from the training domains $\mathcal{E}_{train}$, as $S = \bigcup_{e \in \mathcal{E}_{train}} S_e$,

1.1 Sample Size Within Domains - General Case

In this subsection we assume $\mathcal{L}(h,e) = \mathbb{E}_{(x,y)\sim P_e}[l(h(x),y)]$ for some cost function l.

Theorem 1.1. *Let $\mathcal{H}$, $\mathcal{L}$, γ follow the definitions from Theorem 3.1. Assume that:*

1. *$\mathcal{H}$ has the uniform-convergence property with respect to $\mathcal{L}, \gamma$ according to definition 3 with sample size $m(\delta, \epsilon)$.*
2. *$\mathcal{H}$ has also the uniform-convergence property according to the classical definition (in a single domain scenario) with sample size $n(\delta, \epsilon)$.*

For each $\epsilon_1, \epsilon_2, \delta > 0$, if $|\mathcal{E}_{train}| \geq m(\frac{\delta}{2}, \epsilon_2) := m$, and $\forall e \in \mathcal{E}_{train}$ $|S_e| \geq n(\frac{\delta}{2m}, \epsilon_1) := n$, than with probability higher than $1-\delta$, and regardless of the distribution D over $\mathcal{E}$ and all the distributions P_e for $e \in \mathcal{E}_{train}$, it holds that:

$$\forall h \in \mathcal{H} \quad [\forall_{e \in \mathcal{E}_{train}} \frac{1}{|S_e|} \sum_{i \in S_e} l(h(X_i), y_i) \leq \gamma] \implies \mathop{\mathbb{E}}_{e \sim D}[\mathbb{1}_{\mathcal{L},\gamma+\epsilon_1}(h,e)] \leq \epsilon_2.$$

Proof. let $\epsilon, \delta > 0$ be positive numbers, and assume existences of sample $S = \bigcup_{e \in \mathcal{E}_{train}} S_e$ such that $|\mathcal{E}_{train}| \geq m$ and for each $e \in \mathcal{E}_{train}$ $|S_e| \geq n$.
From the uniform-convergence of $\mathcal{H}$ in the classical, single-domain, sense, we know for each domain $e \in \mathcal{E}_{train}$ that with probability at least $1 - \frac{\delta}{2m}$ it holds that:

$$\forall h \in \mathcal{H} \quad \mathcal{L}(h,e) \leq \frac{1}{|S_e|} \sum_{i \in S_e} l(h(X_i), y_i) + \epsilon_1$$

And so,

$$\forall h \in \mathcal{H} \quad \frac{1}{|S_e|} \sum_{i \in S_e} l(h(X_i), y_i) \leq \gamma \implies \mathcal{L}(h,e) \leq \gamma + \epsilon_1 \implies \mathbb{1}_{\mathcal{L},\gamma+\epsilon_1}(h,e) = 0$$

This is true for each $e \in \mathcal{E}_{train}$, therefore with probability at leat $1 - \frac{\delta}{2}$ it is true for all training domains at once:

$$\forall h \in \mathcal{H} \quad [\forall_{e \in \mathcal{E}_{train}} \frac{1}{|S_e|} \sum_{i \in S_e} l(h(X_i), y_i) \leq \gamma] \Longrightarrow \frac{1}{|\mathcal{E}_{train}|} \sum_{e \in \mathcal{E}_{train}} \mathbb{1}_{\mathcal{L}, \gamma + \epsilon_1}(h, e) = 0$$

From the uniform-convergence property of $\mathcal{H}$ in the OOD sense, we know that with probability at least $1 - \frac{\delta}{2}$ it holds that

$$\forall h \in \mathcal{H} \quad \left| \mathop{\mathbb{E}}_{e \sim D}[\mathbb{1}_{\mathcal{L}, \gamma + \epsilon_1}(h, e)] - \frac{1}{|\mathcal{E}_{train}|} \sum_{e \in \mathcal{E}_{train}} \mathbb{1}_{\mathcal{L}, \gamma + \epsilon_1}(h, e) \right| \leq \epsilon_2.$$

And it the case of $\frac{1}{|\mathcal{E}_{train}|} \sum_{e \in \mathcal{E}_{train}} \mathbb{1}_{\mathcal{L}, \gamma + \epsilon_1}(h, e) = 0$ we get:

$$\forall h \in \mathcal{H} \quad \left| \mathop{\mathbb{E}}_{e \sim D}[\mathbb{1}_{\mathcal{L}, \gamma + \epsilon_1}(h, e)] \right| \leq \epsilon_2.$$

Overall, we have shown that with probability at least $1 - \delta$:

$$\forall h \in \mathcal{H} \quad \forall_{e \in \mathcal{E}_{train}} \frac{1}{|S_e|} \sum_{i \in S_e} l(h(X_i), y_i) \leq \gamma \Longrightarrow \left| \mathop{\mathbb{E}}_{e \sim D}[\mathbb{1}_{\mathcal{L}, \gamma + \epsilon_1}(h, e)] \right| \leq \epsilon_2.$$

□

1.2 Sample Size Within Domains - Recall Loss

Now we focus on $\mathcal{L}_{recall}$. We start by reminding its definition:

$$\mathcal{L}_{recall}(h, e) = \max_{y \in \mathcal{Y}} P_e[y \notin h(X) | Y = y].$$

Now we also assume that $\mathcal{H}$ is an hypothesis set of set-prediction hypotheses, where each $h \in \mathcal{H}$ can be decomposed to $|\mathcal{Y}|$ binary classifiers h_y as presented in the paper. We assume also that the h_y binary classifiers come from some $\mathcal{H}^*$ hypothesis set.

The following result differs from that of Section 1.1 mainly because $\mathcal{L}_{recall}$ is not an expectation over some other loss l. The result we derive for $\mathcal{L}_{recall}$ is almost the same as in the previous section, with only one change: Instead of requiring a sample size of $n(\frac{\delta}{2m}, \epsilon_1)$ at each training domain, we need to require a sample size of $n(\frac{\delta}{2m|\mathcal{Y}|}, \epsilon_1)$ for each training domain and each label $y \in \mathcal{Y}$. For completeness we present here the full result for $\mathcal{L}_{recall}$ and provide a full proof for it.

We add a single notation to this section:

$$S_{e,y} = \{i \in S_e \ : \ Y_i = y\}$$

Theorem 1.2. *Let $\mathcal{H}$, $\mathcal{L}_{recall}$, γ follow the definitions from the paper . Assume that:*

1. *$\mathcal{H}$ has the uniform-convergence property with respect to $\mathcal{L}_{recall}, \gamma$ according to definition 3 with sample size $m(\delta, \epsilon)$.*
2. *$\mathcal{H}^*$ has the uniform-convergence property according to the classical definition (in a single domain scenario) with sample size $n(\delta, \epsilon)$.*

For each $\epsilon_1, \epsilon_2, \delta > 0$, if $|\mathcal{E}_{train}| \geq m(\frac{\delta}{2}, \epsilon_2) := m$, and $\forall e \in \mathcal{E}_{train} \forall y \in \mathcal{Y}$ $|S_{e,y}| \geq n(\frac{\delta}{2m|\mathcal{Y}|}, \epsilon_1) := n$, than with probability higher than $1-\delta$, and regardless of the distribution D over $\mathcal{E}$ and all the distributions P_e for $e \in \mathcal{E}_{train}$, it holds that:

$$\forall h \in \mathcal{H} \quad \left[\forall_{e \in \mathcal{E}_{train}} \forall_{y \in \mathcal{Y}} \frac{1}{|S_{e,y}|} \sum_{i \in S_{e,y}} [1 - h_y(X_i)] \ \leq \gamma\right] \Longrightarrow \mathop{\mathbb{E}}_{e \sim D}[\mathbb{1}_{\mathcal{L}_{recall}, \gamma+\epsilon_1}(h, e)] \leq \epsilon_2.$$

Proof. The main difference in this proof will be to show that with high probability

$$\forall h \in \mathcal{H} \quad \mathcal{L}_{recall}(h, e) \leq \max_{y \in \mathcal{Y}} \frac{1}{|S_{e,y}|} \sum_{i \in S_{e,y}} [1 - h_y(X_i)] + \epsilon_1$$

For competness we provide below the full proof.

let $\epsilon, \delta > 0$ be positive numbers, and assume existences of sample $S = \bigcup_{e \in \mathcal{E}_{train}} S_e$ such that $|\mathcal{E}_{train}| \geq m$ and for each $e \in \mathcal{E}_{train}$ $|S_e| \geq n$.
From the uniform-convergence of $\mathcal{H}^*$ in the classical, single-domain, sense, we know for each domain $e \in \mathcal{E}_{train}$ and for each $y \in \mathcal{Y}$ that with probability at least $1 - \frac{\delta}{2m|\mathcal{Y}|}$ it holds that:

$$\forall h \in \mathcal{H}^* \quad P_e[h(X) \neq 1 | Y = y] = \mathbb{E}_e[1 - h(X) | Y = y] \leq \frac{1}{|S_{e,y}|} \sum_{i \in S_{e,y}} [1 - h(X_i)] + \epsilon_1$$

This is true for each y separately, so with probability at least $1 - \frac{\delta}{2m}$ this is true for all y at once:

$$\forall h \in \mathcal{H} \quad \mathcal{L}_{recall}(h, e) = \max_{y \in \mathcal{Y}} P_e[y \notin h(X) | Y = y] \leq \max_{y \in \mathcal{Y}} \frac{1}{|S_{e,y}|} \sum_{i \in S_{e,y}} [1 - h_y(X_i)] + \epsilon_1$$

And so,

$$\forall h \in \mathcal{H} \quad \max_{y \in \mathcal{Y}} \frac{1}{|S_{e,y}|} \sum_{i \in S_{e,y}} [1 - h_y(X_i)] \leq \gamma \Longrightarrow \mathcal{L}_{recall}(h, e) \leq \gamma + \epsilon_1 \Longrightarrow$$

$$\Longrightarrow \mathbb{1}_{\mathcal{L}_{recall}, \gamma+\epsilon_1}(h, e) = 0$$

This is true for each $e \in \mathcal{E}_{train}$, therefore with probability at leat $1 - \frac{\delta}{2}$ it is true for all training domains at once:

$$\forall h \in \mathcal{H} \quad [\forall_{e \in \mathcal{E}_{train}} \max_{y \in \mathcal{Y}} \frac{1}{|S_{e,y}|} \sum_{i \in S_{e,y}} [1 - h_y(X_i)] \leq \gamma] \Longrightarrow \frac{1}{|\mathcal{E}_{train}|} \sum_{e \in \mathcal{E}_{train}} \mathbb{1}_{\mathcal{L}_{recall}, \gamma + \epsilon_1}(h, e) = 0$$

From the uniform-convergence property of $\mathcal{H}$ in the OOD sense, we know that with probability at least $1 - \frac{\delta}{2}$ it holds that

$$\forall h \in \mathcal{H} \quad \left| \mathop{\mathbb{E}}_{e \sim D}[\mathbb{1}_{\mathcal{L}_{recall}, \gamma + \epsilon_1}(h, e)] - \frac{1}{|\mathcal{E}_{train}|} \sum_{e \in \mathcal{E}_{train}} \mathbb{1}_{\mathcal{L}_{recall}, \gamma + \epsilon_1}(h, e) \right| \leq \epsilon_2.$$

And it the case of $\frac{1}{|\mathcal{E}_{train}|} \sum_{e \in \mathcal{E}_{train}} \mathbb{1}_{\mathcal{L}_{recall}, \gamma + \epsilon_1}(h, e) = 0$ we get:

$$\forall h \in \mathcal{H} \quad \left| \mathop{\mathbb{E}}_{e \sim D}[\mathbb{1}_{\mathcal{L}_{recall}, \gamma + \epsilon_1}(h, e)] \right| \leq \epsilon_2.$$

Overall, we have shown that with probability at least $1 - \delta$:

$$\forall h \in \mathcal{H} \quad \forall_{e \in \mathcal{E}_{train}} \forall_{y \in \mathcal{Y}} \frac{1}{|S_{e,y}|} \sum_{i \in S_{e,y}} [1 - h_y(X_i)] \ \leq \gamma \Longrightarrow \mathop{\mathbb{E}}_{e \sim D}[\mathbb{1}_{\mathcal{L}_{recall}, \gamma + \epsilon_1}(h, e)] \leq \epsilon_2.$$

□

1.3 Sample Size Within Domains - Recall Loss With Linear Hypotheses

finally, we show the sample complexity for $\mathcal{L}_{recall}$ when $\mathcal{H}$ is the set of linear hypotheses.

Theorem 1.3. *Let $\mathcal{H}$ be the hypothesis set of linesr set predictors in $\mathbb{R}^d$. Assume domains are restricted to being Conditionally Gaussian as described in Theorem 3.4.*
For each $\epsilon_1, \epsilon_2, \delta > 0$, if $|\mathcal{E}_{train}| \geq \Theta(\frac{|\mathcal{Y}|(d + log(2|\mathcal{Y}|/\delta))}{\epsilon_2^2}) := m$, and $\forall e \in \mathcal{E}_{train} \forall y \in \mathcal{Y}$ $|S_{e,y}| \geq \Theta(\frac{d + log(2m|\mathcal{Y}|/\delta))}{\epsilon_1^2})$, than with probability higher than $1 - \delta$, and regardless of the distribution D over $\mathcal{E}$ and all the distributions P_e for $e \in \mathcal{E}_{train}$, it holds that:

$$\forall h \in \mathcal{H} \quad \left[\forall_{e \in \mathcal{E}_{train}} \forall_{y \in \mathcal{Y}} \frac{1}{|S_{e,y}|} \sum_{i \in S_{e,y}} [1 - h_y(X_i)] \ \leq \gamma\right] \Longrightarrow \mathop{\mathbb{E}}_{e \sim D}[\mathbb{1}_{\mathcal{L}_{recall}, \gamma + \epsilon_1}(h, e)] \leq \epsilon_2.$$

Proof. In the classical, single-domain context, linear hypotheses have $VC - dim = d + 1$, and they hold the uniform-convergence property with $n(\delta, \epsilon) = \Theta(\frac{d + log(1/\delta))}{\epsilon^2})$.

Following the exact same steps from the proof of the previous section, we can derive that with probability at leat $1 - \frac{\delta}{2}$:

$$\forall h \in \mathcal{H} \quad [\forall_{e \in \mathcal{E}_{train}} \max_{y \in \mathcal{Y}} \frac{1}{|S_{e,y}|} \sum_{i \in S_{e,y}} [1 - h_y(X_i)] \leq \gamma] \Longrightarrow \forall e \in \mathcal{E}_{train} \mathbb{1}_{\mathcal{L}_{recall}, \gamma+\epsilon_1}(h, e) = 0$$

Now, from Theorem 3.4 we know that with probability at leat $1 - \frac{\delta}{2}$:

$$\forall h \in \mathcal{H} \quad \forall e \in \mathcal{E}_{train} \mathbb{1}_{\mathcal{L}_{recall}, \gamma+\epsilon_1}(h, e) = 0 \Longrightarrow \mathop{\mathbb{E}}_{e \sim D} [\mathbb{1}_{\mathcal{L}_{recall}, \gamma+\epsilon_1}(h, e)] \leq \epsilon_2.$$

Together, we get that with probability at least $1 - \delta$:

$$\forall h \in \mathcal{H} \quad [\forall_{e \in \mathcal{E}_{train}} \max_{y \in \mathcal{Y}} \frac{1}{|S_{e,y}|} \sum_{i \in S_{e,y}} [1 - h_y(X_i)] \leq \gamma] \Longrightarrow \mathop{\mathbb{E}}_{e \sim D} [\mathbb{1}_{\mathcal{L}_{recall}, \gamma+\epsilon_1}(h, e)] \leq \epsilon_2.$$

$\square$