The No-Free-Lunch Theorem

Prof. Dan A. Simovici

UMB

1. Preliminaries
2. The No-Free-Lunch Theorem

Reminder

- If K is event such that $P(K) = p$, $\mathbf{1}_K$ is a random variable

$$\mathbf{1}_K = \begin{cases} 1 & \text{if } K \text{takes place} \\ 0 & \text{otherwise.} \end{cases}$$

- If $P(K) = p$, then

$$\mathbf{1}_K : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

  and $E(\mathbf{1}_K) = p$.
- If X is a random variable

$$X : \begin{pmatrix} x_1 & \cdots & x_n \\ p_1 & \cdots & p)n \end{pmatrix},$$

  then $X = \sum_{i=1}^n x_i \mathbf{1}_{X=x_i}$, where

$$\mathbf{1}_{X=x_i} : \begin{pmatrix} 0 & 1 \\ 1-p_i & p_i \end{pmatrix}.$$

First Lemma

Lemma

Let Z be a random variable that takes values in $[0,1]$ such that $E[Z] = \mu$. Then, for every $a \in (0,1)$ we have

$$P(Z > 1-a) \geqslant \frac{\mu - (1-a)}{a} \text{ and } P(Z > a) \geqslant \frac{\mu - a}{1-a} \geqslant \mu - a.$$

Proof

The random variable $Y = 1 - Z$ is non-negative with $E(Y) = 1 - E(Z) = 1 - \mu$. By Markov's inequality:

$$P(Z \leqslant 1 - a) = P(1 - Z \geqslant a) = P(Y \geqslant a) \leqslant \frac{E(Y)}{a} = \frac{1 - \mu}{a}.$$

Therefore,

$$P(Z > 1 - a) \geqslant 1 - \frac{1 - \mu}{a} = \frac{a + \mu - 1}{a} = \frac{\mu - (1 - a)}{a}.$$

By replacing a by $1 - a$ we have:

$$P(Z > a) \geqslant \frac{\mu - a}{1 - a} \geqslant \mu - a.$$

Second Lemma

Lemma

Let θ *be a random variable that ranges in the interval* $[0,1]$ *such that* $E(\theta) \geqslant \frac{1}{4}$. *We have*

$$P\left(\theta > \frac{1}{8}\right) \geqslant \frac{1}{7}.$$

Proof: From the second inequality of the previous lemma it follows that

$$P(\theta > a) \geqslant \frac{E(\theta) - a}{1 - a}.$$

By substituting $a = \frac{1}{8}$ we obtain:

$$P(\theta > \frac{1}{8}) \geqslant \frac{\frac{1}{4} - \frac{1}{8}}{1 - \frac{1}{8}} = \frac{1}{7}.$$

- A learning task is defined by an unknown probability distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$.
- The goal of the learner is to find (to learn) a hypothesis $h : \mathcal{X} \longrightarrow \mathcal{Y}$ such that its risk $L_{\mathcal{D}}(h)$ is sufficiently small.
- The choice of a hypothesis class $\mathcal{H}$ reflects some prior knowlege that the learner has about the task: a belief that a member of $\mathcal{H}$ is a low-error model for the task.
- Fundamental Question: There exist a universal learner $\mathcal{A}$ and a training set size m such that for every distribution $\mathcal{D}$, if $\mathcal{A}$ receives m iid examples from $\mathcal{D}$, there is a high probability that $\mathcal{A}$ will produce h with a low risk?

- The No-Free-Lunch (NFL) Theorem stipulates that a universal learner does no exist!
- A learner fails if, upon receiving a sequence of iid examples from a distribution $\mathcal{D}$, its output hypothesis is likely to have a large loss (say, larger than 0.3), whereas for the same distribution there exists another learner that will output a hypothesis with a small risk.
- More precise statment: for every binary prediction task and learner, there exists a distribution $\mathcal{D}$ for which the learning task fails.
- No learner can succeed on all learning tasks: every learner has tasks on which it fails whereas other learners succeed.

Recall 0/1 Loss Function

The random variable z ranges over $\mathcal{X} \times \mathcal{Y}$ and the loss function is

$$\ell_{0-1}(h,(x,y)) = \begin{cases} 0 & \text{if } h(x) = y, \\ 1 & \text{if } h(x) \neq y. \end{cases}$$

The NFL Theorem

Theorem

Let $\mathcal{A}$ be any learning algorithm for the task of binary classification with respect to the $0/1$-loss function over a domain $\mathcal{X}$. Let $m < \frac{|\mathcal{X}|}{2}$ be a number representing a training set size.

There exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$ such that:

- *there exists a function $f : \mathcal{X} \longrightarrow \{0, 1\}$ with $L_{\mathcal{D}}(f) = 0$;*
- *with probability at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(\mathcal{A}(S)) \geqslant 1/8$.*

Proof

Since $m < \frac{|\mathcal{X}|}{2}$, there exists a subset C of $\mathcal{X}$ of size $2m$.
Intuition of the proof: any algorithm that observes only half of the instances of C has no information of what should be the labels of the other half. Therefore, there exists a target function f which would contradict the labels that $\mathcal{A}(S)$ predicts on the unobserved instances of C.

- There are $T = 2^{2m}$ possible functions from C to $\{0,1\}$: $f_1, \ldots, f_T$.
- For each f_i let $\mathcal{D}_i$ be the distribution over

$$C \times \{0,1\} = \{(x_1,0),(x_1,1),\ldots,(x_{2m},0),(x_{2m},1)\}$$

given by

$$\mathcal{D}_i(\{(x,y)\}) = \begin{cases} \frac{1}{|C|} & \text{if } y = f_i(x) \\ 0 & \text{otherwise.} \end{cases}$$

The probability to choose a pair (x,y) is $\frac{1}{|C|}$ if y is the true label according to f_i and 0, otherwise (if $y \neq f_i(x)$). Clearly $L_{\mathcal{D}_i}(f_i) = 0$.

Intuition

Let $m = 3$, $C = \{x_1, x_2, x_3, x_4, x_5, x_6\}$.
Suppose that

$$\begin{array}{ccc} f_i(x_1) = 1 & f_i(x_2) = 0 & f_i(x_3) = 1 \\ f_i(x_4) = 1 & f_i(x_5) = 1 & f_i(x_6) = 0, \end{array}$$

The distribution $\mathcal{D}_i$ is:

$$\begin{array}{cccccc} (x_1, 0) & (x_2, 0) & (x_3, 0) & (x_4, 0) & (x_5, 0) & (x_6, 0) \\ 0 & \frac{1}{6} & 0 & 0 & 0 & \frac{1}{6} \\ (x_1, 1) & (x_2, 1) & (x_3, 1) & (x_4, 1) & (x_5, 1) & (x_6, 1) \\ \frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0. \end{array}$$

Clearly, we have:

$$L_{\mathcal{D}_i}(f_i) = P(\{(x, y) \mid f(x) \neq y\}) = 0.$$

Claim (*):
For every algorithm $\mathcal{A}$ that receives a training set of m examples from $C \times \{0,1\}$ and returns a function $\mathcal{A}(S) : C \longrightarrow \{0,1\}$ we have:

$$\max_{1 \leqslant i \leqslant |T|} E_{S \sim \mathcal{D}^m}(L_{D_i}(A(S))) \geqslant \frac{1}{4}.$$

This means that for every $\mathcal{A}'$ that receives a training set of m examples from $\mathcal{X} \times \{0,1\}$ there exists $f : \mathcal{X} \longrightarrow \{0,1\}$ and a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that $L_{\mathcal{D}}(f) = 0$ and $E_{S \sim \mathcal{D}^m}(L_{cald}(\mathcal{A}'(S))) \geqslant \frac{1}{4}$.

- There are $k = (2m)^m$ possible sequences $S_1, \ldots, S_k$ of m examples from C.
- If $S_j = (x_1, \ldots, x_m)$, the sequence labeled by a function f_i is denoted by $S_j^i = ((x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m)))$.
- If the distribution is $\mathcal{D}_i$, then the possible training sets that $\mathcal{A}$ can receive are $S_1^i, \ldots, S_k^i$ and all these training sets have the same probability of being sampled. Therefore,

$$E_{S \sim \mathcal{D}^m}(L_{\mathcal{D}_i}(\mathcal{A}(S)) = \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i)).$$

Recall that there are $T = 2^{2m}$ possible functions from C to $\{0, 1\}$: $f_1, \ldots, f_T$, so $1 \leqslant i \leqslant T$, where i the superscript of S_j^i reflecting the labeling function f_i.
We have:

$$\begin{aligned}
& \max_{1 \leqslant i \leqslant T} \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i)) \\
\geqslant \quad & \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i)) \\
= \quad & \frac{1}{k} \sum_{j=1}^{k} \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i)) \\
\geqslant \quad & \min_{1 \leqslant j \leqslant k} \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i)).
\end{aligned}$$

Fix some j and let $S_j = (x_1, \ldots, x_m)$. Let $v_1, \ldots, v_p$ be the examples in C that do not appear in S_j. Clearly $p \geqslant m$. Therefore, for each $h : C \longrightarrow \{0, 1\}$ and every i we have

$$\begin{aligned} L_{\mathcal{D}_i}(h) &= \frac{1}{2m} \sum_{x \in C} \mathbf{1}_{h(x) \neq f_i(x)} \geqslant \frac{1}{2m} \sum_{r=1}^{p} \mathbf{1}_{h(v_r) \neq f_i(v_r)} \\ &\geqslant \frac{1}{2p} \sum_{r=1}^{p} \mathbf{1}_{h(v_r) \neq f_i(v_r)}. \end{aligned}$$

Hence,

$$\begin{aligned} \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i)) &\geqslant \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2p} \sum_{r=1}^{p} \mathbf{1}_{\mathcal{A}(S_j^i)(v_r) \neq f_i(v_r)} \\ &= \frac{1}{2p} \sum_{r=1}^{p} \frac{1}{T} \sum_{i=1}^{T} \mathbf{1}_{\mathcal{A}(S_j^i)(v_r) \neq f_i(v_r)} \\ &\geqslant \frac{1}{2} \min_{1 \leqslant t \leqslant p} \frac{1}{T} \sum_{i=1}^{T} \mathbf{1}_{\mathcal{A}(S_j^i)(v_r) \neq f_i(v_r)}. \end{aligned}$$

Fix some r, $1 \leqslant r \leqslant p$. We can partition all functions in $\{f_1, \ldots, f_p\}$ into $T/2$ disjoint sets $\{f_i, f_{i'}\}$ such that we have

$$f_i(c) \neq f_{i'}(c) \text{ if and only if } c = v_r.$$

Since for a set $\{f_i, f_{i'}\}$ we must have $S_j^i = S_j^{i'}$, it follows that

$$1_{A(S_j^i)(v_r) \neq f_i(v_r)} + 1_{A(S_j^{i'})(v_r) \neq f_{i'}(v_r)} = 1,$$

which implies

$$\frac{1}{T} \sum_{i=1}^{T} 1_{A(S_j^i(v_r) \neq f_i(v_r)} = \frac{1}{2}.$$

Since

$$\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i) \geqslant \frac{1}{2}\min_{1\leqslant t\leqslant p}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{\mathcal{A}(S_j^i)(v_r)\neq f_i(v_r)}$$

and

$$\frac{1}{T}\sum_{i=1}^{T} 1_{A(S_j^i(v_r)\neq f_i(v_r)} = \frac{1}{2},$$

we have

$$\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i) \geqslant \frac{1}{4}.$$

Thus,

$$\max_{1 \leqslant i \leqslant T} \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i)) \geqslant \min_{1 \leqslant j \leqslant k} \frac{1}{T} \sum_{i=1}^{T} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i))$$

implies

$$\max_{1 \leqslant i \leqslant T} \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i)) \geqslant \frac{1}{4}.$$

We combined

$$
\begin{aligned}
\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i)) &\geqslant \frac{1}{2}\min_{1\leqslant t\leqslant p}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{\mathcal{A}(S_j^i)(v_r)\neq f_i(v_r)}\\
\max_{1\leqslant i\leqslant T}\frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i)) &\geqslant \min_{1\leqslant j\leqslant k}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i))\\
E_{S\sim\mathcal{D}^m}(L_{\mathcal{D}_i}(\mathcal{A}(S))) &= \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(\mathcal{A}(S_j^i))\\
\frac{1}{T}\sum_{i=1}^{T} 1_{A(S_j^i)(v_r)\neq f_i(v_r)} &= \frac{1}{2}
\end{aligned}
$$

to obtain:

$$
\max_{1\leqslant i\leqslant T} E_{S\sim\mathcal{D}_i^m}(L_{\mathcal{D}_i}(\mathcal{A}(S)) \geqslant \frac{1}{4}.
$$

Thus, the Claim (*) is justified.
This means that for every algorithm $\mathcal{A}'$ that receives a training set of m examples from $\mathcal{X} \times \{0,1\}$ there exists a function $f : \mathcal{X} \longrightarrow \{0,1\}$ and a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that $L_{\mathcal{D}}(f) = 0$ and

$$E_{S \sim \mathcal{D}^m}(L_{\mathcal{D}}(\mathcal{A}'(S))) \geqslant \frac{1}{4}.$$

By the second Lemma this implies:

$$P(L_{\mathcal{D}}(\mathcal{A}'(S)) \geqslant \frac{1}{8}) \geqslant \frac{1}{7}.$$