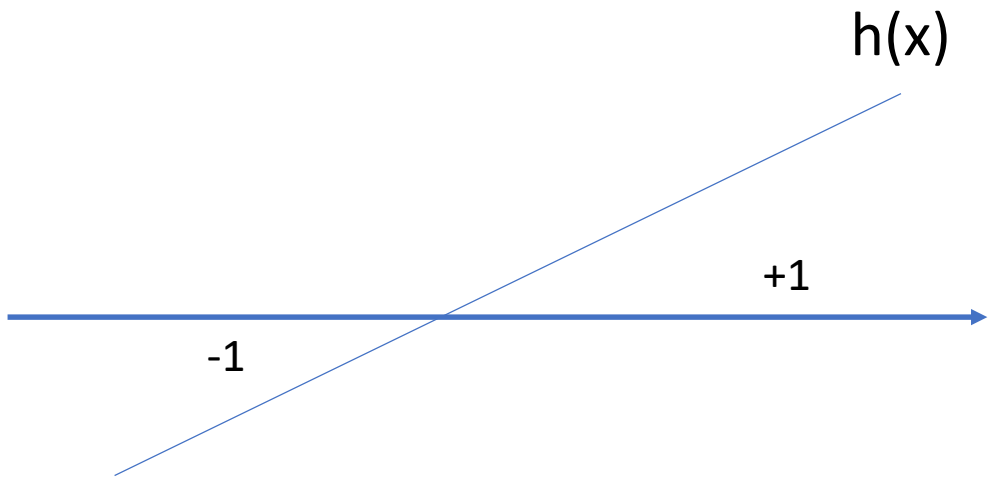
Non-Uniform Learnability and Structural Risk Minimization

Example

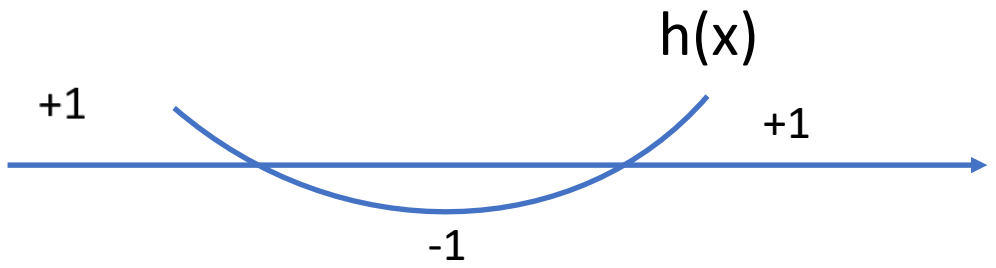
- $x \in \mathbb{R}$ (blood pressure of a patient)
- h(x) = sign(ax+b) (will have heart attack in an hour?)
- H = {h|h(x), a, b}



ERM will give us the best linear classifier, but its accuracy may not be good enough

Example

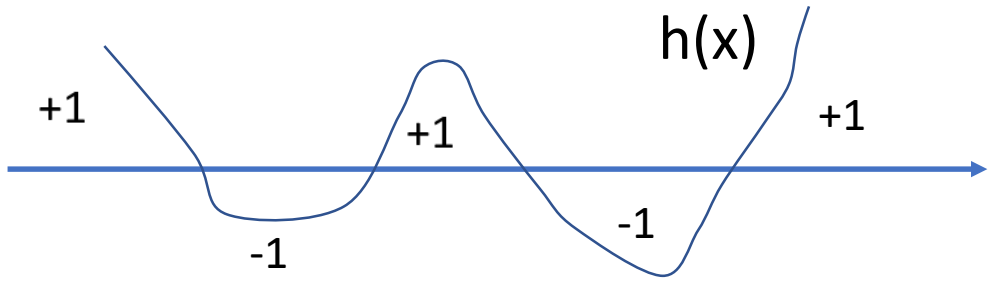
- $x \in \mathbb{R}$ (blood pressure of a patient)
- $h(x) = \text{sign}(ax^2 + bx + c)$ (will have heart attack in an hour?)
- $H = \{h | h(x), a, b, c\}$



ERM will give us the best linear classifier, but its accuracy may still not be good enough

Example

- $x \in \mathbb{R}$ (blood pressure of a patient)
- h(x) = sign(d-degree polynomial(x)) (will have heart attack in an hour?)
- H = {h|h(x), a, b, c, d}



ERM will give us the best linear classifier, but its accuracy may still not be good enough!

A More General Hypothesis Class

- $H = \cup \{H_{linear}, H_{quadratic}, H_{cubic}, H_{poly\text{-}d}, \ldots H_{poly\text{-}\infty}\}$
- We keep the hypothesis class general and will choose a suitable hypothesis depending on the problem.
- VC dimension of H is ∞ since it is essentially the class of all possible binary functions on the real line
- Can we still learn?

Recap: Agnostic PAC Learnability

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a set Z and a loss function $l : Z \times \mathcal{H} \to \mathbb{R}_+$ if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A with the following property:

- for every $\epsilon, \delta \in (0,1)$
- for every distribution $\mathcal{D}$ over Z
- when running A on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. samples generated by $\mathcal{D}$
- A returns a hypothesis $h \in \mathcal{H}$ such that with probability at least $1 - \delta$

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

Uniform Learnability

- Uniform sample complexity requirement for all h ∈ H

$$m \geq m_{\mathcal{H}}(\epsilon, \delta)$$

Non-Uniform Learnability

- Allows varying sample complexity requirement for different h ∈ H

$$m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$$

- Intuition:
- We will need relatively less number of samples to learn simpler classifiers
- More complex the classifier - more data we need.

- Cost increases as we try to match sophisticated classifiers!

Definition: Non-Uniform Learnability

A hypothesis class $\mathcal{H}$ is non-uniformly learnable if there exists a learning algorithm A and a function $m_{\mathcal{H}}^{NUL} : (0,1)^2 \times \mathcal{H} \to \mathbb{N}$ such that

- for every $\epsilon, \delta \in (0,1)$
- for every $h \in \mathcal{H}$
- when running A on $m \geq m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$ i.i.d. samples
- then for every distribution $\mathcal{D}$ over Z
- it holds that for with probability at least $1 - \delta$ over the choice of $D \sim \mathcal{D}^m$

$$L_{\mathcal{D}}(A(D)) \leq L_{\mathcal{D}}(h) + \epsilon$$

NUL Requires Existence of Learner A

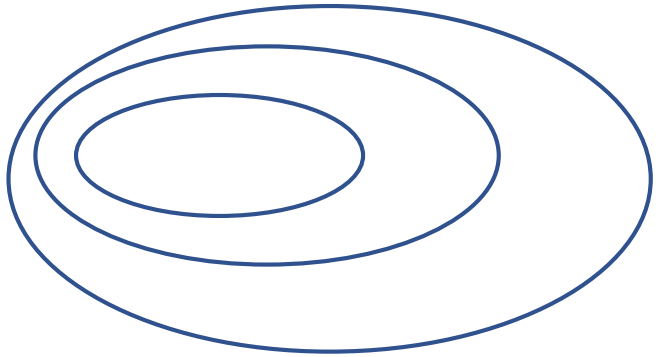
Given a data set, A will, with high probability, deliver a competitive hypothesis; that is, competitive with those hypotheses whose sample complexity is less than $|D|$.

Example of NUL Hypothesis Class

- Countable union of uniformly learnable hypothesis classes
  - May not be uniformly learnable
  - May be non-uniformly learnable

A Structure on the Union Class

- that $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$
- and a weight function $w : \mathbb{N} \to [0, 1]$



Background Knowledge of Structure

- that $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$
- and a weight function $w : \mathbb{N} \to [0, 1]$

Both can be seen as a form of background knowledge

- the choice of $\mathcal{H}$ itself is already background knowledge, putting structure to it even more so
- all the more since w allows us to specify where in $\mathcal{H}$ we expect it to be likely to find the model ($w(n)$ high, chance of $\mathcal{H}_n$ high)

An Analogy of Structured Hypothesis Class

- We try to diagnose a disease by consulting a doctor
- Doctors have varying degree of expertise
  - Doctors belongs to groups H_n depending on their highest degree n
  - Expertise/Capacity of H_n increases with n (as encoded in weight w_n)

- Cost of consulting (sample complexity) a doctor increases with her expertise
- We want to choose the best doctor for the diagnosis but we prefer to pay a small fee

Non-Uniform Learnability of the Union Class

A hypothesis class $\mathcal{H}$ of binary classifiers is non-uniformly learnable iff it is the countable union of agnostic PAC learnable hypothesis classes.

Let $\mathcal{H}$ be a hypothesis class that can be written as a countable union $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$, where for all n, $VC(\mathcal{H}_n) < \infty$, then $\mathcal{H}$ is non-uniformly learnable.

Theorem

A hypothesis class $\mathcal{H}$ of binary classifiers is non-uniformly learnable iff it is the countable union of agnostic PAC learnable hypothesis classes.

Proving - Only If

Let $\mathcal{H}$ be non-uniformly learnable. That means that we have a function $m_{\mathcal{H}}^{NUL} : (0,1)^2 \times \mathcal{H} \to \mathbb{N}$ to compute sample sizes.

- for a given ϵ_0, δ_0 define for every $n \in \mathbb{N}$

$$\mathcal{H}_n = \{h \in \mathcal{H} \mid m_{\mathcal{H}}^{NUL}(\epsilon_0, \delta_0, h) \leq n\}$$

- clearly, for every ϵ_0 and δ_0 we have that

$$\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$$

- Moreover, for every $h \in \mathcal{H}_n$ we know that with probability of at least $1 - \delta_0$ over $D \sim \mathcal{D}^n$ we have $L_{\mathcal{D}}(A(D)) \leq L_{\mathcal{D}}(h) + \epsilon_0$.
- since this holds *uniformly* for all $h \in \mathcal{H}_n$
- we have that H_n is agnostic PAC learnable

Proving - If

- If the individual classes are uniformly learnable the union is non-uniformly learnable

Proof Outline for Uniform Convergence

1. Show that all samples of size $m > m_H(\varepsilon, \delta)$ are ε-representative

$$\forall h \in \mathcal{H} : |L_D(h) - L_{\mathcal{D}}(h)| \leq \epsilon$$

2. ERM succeeds on a ε-representative sample to attain -

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

Proof Outline for Non-Uniform Convergence

1. Show that all samples of size $m > m_{H_n}(\varepsilon, \delta, H_n)$ are ε_n-representative

$$\forall h \in \mathcal{H} : |L_D(h) - L_{\mathcal{D}}(h)| \leq \epsilon$$

2. SRM succeeds on a ε-representative sample to attain -

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$$

The ϵ_n Function

We assume that $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$

- and that each $\mathcal{H}_n$ has the uniform convergence property

Now define the function $\epsilon_n : \mathbb{N} \times (0,1) \to (0,1)$ by

$$\epsilon_n(m, \delta) = \min\{\epsilon \in (0,1) \mid m_{\mathcal{H}_n}^{UC}(\epsilon, \delta) \leq m\}$$

That is, given a fixed sample size, we are interested in the smallest possible gap between empirical and true risk. To see this, substitute $\epsilon_n(m, \delta)$ in the definition of uniform convergence, then we get:

For every m and δ with probability of at least $1 - \delta$ over the choice of $D \sim \mathcal{D}^m$ we have

$$\forall h \in \mathcal{H}_n : |L_{\mathcal{D}}(h) - L_D(h)| \leq \epsilon_n(m, \delta)$$

This is the bound we want to extend to all of $\mathcal{H}$

The Weight Function

For that we use the weight function $w : \mathbb{N} \to [0, 1]$. Not any such function will do, it should be a convergent sequence, more precisely we require that

$$\sum_{i=1}^{\infty} w(n) \leq 1$$

In a finite case, this is easy to achieve

- if you have no idea which $\mathcal{H}_n$ is best you can simply choose a uniform distribution

In the countable infinite case you *can not* do that

Bounding Non-Uniform Loss

Let $w : \mathbb{N} \to [0, 1]$ be a function such that $\sum_{i=1}^{\infty} w(n) \leq 1$. Let $\mathcal{H}$ be a hypothesis class that can be written as $\cup_{n \in \mathbb{N}} \mathcal{H}_n$ where each each $\mathcal{H}_n$ has the uniform convergence property. Let $\epsilon_n(m, \delta)$ be as defined before, i.e., $\min\{\epsilon \in (0, 1) \mid m^{UC}_{\mathcal{H}_n}(\epsilon, \delta) \leq m\}$. Then

- for every $\delta \in (0, 1)$ and every distribution $\mathcal{D}$
- with probability of at least $1 - \delta$ over the choice of $D \sim \mathcal{D}^m$

$$\forall n \in \mathbb{N}\ \forall h \in \mathcal{H}_n : |L_{\mathcal{D}}(h) - L_D(h)| \leq \epsilon_n(m, w(n)\delta)$$

Therefore, every $\delta \in (0, 1)$ and every distribution $\mathcal{D}$ with probability of at least $1 - \delta$

$$\forall h \in \mathcal{H} : L_{\mathcal{D}}(h) \leq L_D(h) + \min_{\substack{n \in \mathbb{N} \\ h \in \mathcal{H}_n}} \epsilon_n(m, w(n)\delta)$$

Proof of the Bound

Define for $n \in \mathbb{N}$, $\delta_n = w(n)\delta$. Then we know that if we fix n

- we have with probability at least $1 - \delta_n$ over the choice of $D \sim \mathcal{D}^m$

$$\forall h \in \mathcal{H}_n : |L_{\mathcal{D}}(h) - L_D(h)| \leq \epsilon_n(m, \delta_n)$$

Applying the union bound over $n = 1, 2, \ldots$ then gives us that

- with probability at least

$$1 - \sum_{n \in \mathbb{N}} \delta_n = 1 - \delta \sum_{n \in \mathbb{N}} w(n) \geq 1 - \delta$$

- that

$$\forall n \in \mathbb{N}\ \forall h \in \mathcal{H}_n : |L_{\mathcal{D}}(h) - L_D(h)| \leq \epsilon_n(m, \delta_n)$$

An Upper Bound on Risk

The error you estimate for a $h \in \mathcal{H}$ depends on the $\mathcal{H}_n$, h is a member of. If it is a member of multiple, one should, of course, go for the smallest n:

$$n(h) = \min\{n \mid h \in \mathcal{H}_n\}$$

Then we have

$$L_{\mathcal{D}}(h) \leq L_D(h) + \epsilon_{n(h)}(m, w(n(h))\delta)$$

Learner Algorithm: Structural Risk Minimization

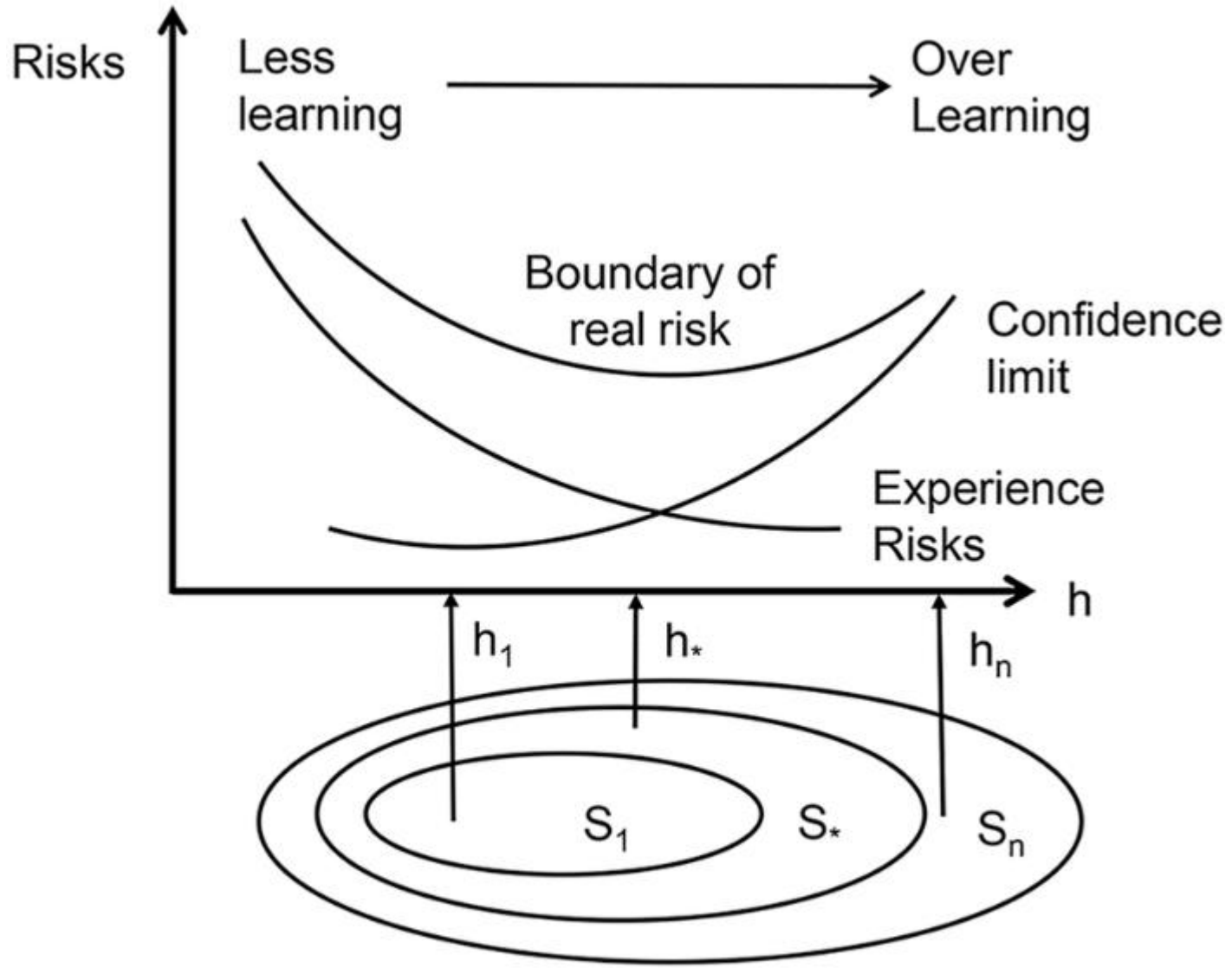
The Structural Risk Minimization Learning Rule is to output

$$h \in \operatorname*{argmin}_{h \in \mathcal{H}} \left[L_D(h) + \epsilon_{n(h)}(m, w(n(h))\delta) \right]$$

So, not just minimal empirical risk, but a balance between

- the empirical risk $L_D(h)$
- and the "class-risk" $\epsilon_{n(h)}(m, w(n(h))\delta)$

SRM



Subsets of function sets :$S_1 \subseteq S_* \subseteq S_n$

Vapnik-Chervonenkis Dimension: $h_1 \leq h_* \leq h_n$

SRM Learning Works

Let $\mathcal{H}$ be a hypothesis class that can be written as $\cup_{n\in\mathbb{N}}\mathcal{H}_n$ where each each $\mathcal{H}_n$ has the uniform convergence property with sample complexity $m^{UC}_{\mathcal{H}_n}$. Let $w(n) = \frac{6}{n^2\pi^2}$. Then

- $\mathcal{H}$ is non-uniformly learnable using the SRM rule with sample complexity

$$m^{NUL}_{\mathcal{H}}(\epsilon, \delta, h) \leq m^{UC}_{\mathcal{H}_{n(h)}}\left(\epsilon/2, \frac{6\delta}{(\pi n(h))^2}\right)$$

Note that

- this theorem does hold far more general than for this specific weight function only

Proof

First of all, note that $\sum_{n \in \mathbb{N}} w(n) = 1$. Next, let A be the SRM learning algorithm with respect to $w(n)$ And for all $h \in \mathcal{H}$, ϵ, and δ, let $m \geq m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon, w(n(h))\delta)$.

- then, with probability at least $1 - \delta$ for the choice of $D \sim \mathcal{D}^m$
- for all $h' \in \mathcal{H}$

$$L_{\mathcal{D}}(h') \leq L_D(h') + \epsilon_{n(h')}(m, w(n(h'))\delta)$$

This holds in particular for hypothesis $A(D)$. By the definition of SRM we get:

$$\begin{aligned} L_{\mathcal{D}}(A(D)) &\leq \min_{h'} \left[L_D(h') + \epsilon_{n(h')}(m, w(n(h'))\delta) \right] \\ &\leq L_D(h) + \epsilon_{n(h)}(m, w(n(h))\delta) \end{aligned}$$

Proof (Contd.)

So, we have that $L_{\mathcal{D}}(A(D)) \leq L_D(h) + \epsilon_{n(h)}(m, w(n(h))\delta)$.

- by definition we have that $m \geq m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon/2, w(n(h))\delta)$ implies that $\epsilon_{n(h)}(m, w(n(h))\delta) \leq \epsilon/2$.
- moreover, because the $\mathcal{H}_n$ have the universal convergence property , we now with probability at least $1 - \delta$:

$$L_D(h) \leq L_{\mathcal{D}}(h) + \epsilon/2$$

That is:

$$\begin{aligned} L_{\mathcal{D}}(A(D)) &\leq L_D(h) + \epsilon_{n(h)}(m, w(n(h))\delta) \\ &\leq L_{\mathcal{D}}(h) + \epsilon/2 + \epsilon/2 \\ &\leq L_{\mathcal{D}}(h) + \epsilon \end{aligned}$$

Examples of SRM

- Depending on how you choose your weight function we get a varierty of SRM algorithms
- Maximum margin classifiers
  - Support Vector Machines

- Minimum Description Length Principle

$$h \in \operatorname*{argmin}_{h \in \mathcal{H}} \left[L_D(h) + \sqrt{\frac{|h| + \log(2/\delta)}{2m}} \right]$$