Agnostic PAC Learning

Notations Recap

- X – Input
- Y- Output/Label
- *D* – Data distribution of X
- *f*(x) – target/labelling function
- Data generation model
    - x ~ *D*
    - *y = f(x)*
- *h(x)* ∈ *H* – Hypothesis

- Any domain
- {0, 1}
- Fixed but unknown
- X → Y

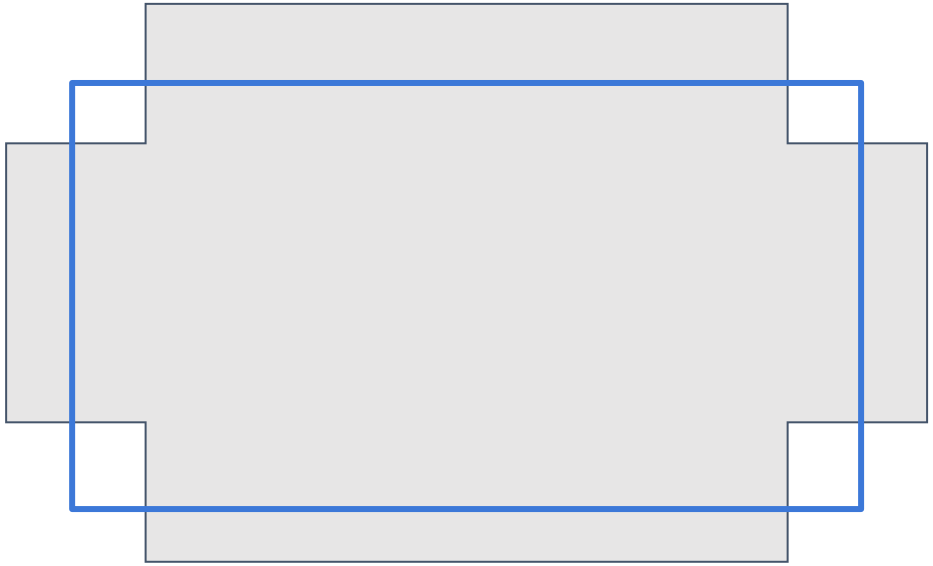
- Realizability assumption: *f*(x) ∈ *H*

Empirical Risk Minimization (ERM)

- A learning paradigm -
- h_S = argmin $_{h \in H} L_S(h)$
- Under the realizability assumption $f \in H$
  - $h_S = {}_{h \in H} L_S(h) = 0$

Relaxation of the Realizability Assumption

- $f(x) \notin H$
- Approximating a complex target function with a simpler hypothesis class
- Approximation error



Relaxation of Realizability Assumption

- No *f(x)* exists

- Non-zero Bayes error - overlapping classes

More General Data Generation Process

- Both input and output are samples from a distribution D
  - $Z = (X, y) \sim D$
  - D is a distribution over X x y

- Allows for overlapping classes
- No underlying function $y = f(X)$

More General Loss Function

- $L_D(h) = \mathrm{E}_D(\ell(h(x),y))$

- Expectation over D of random variable $\ell(h(x),y))$
- More general loss function $\ell(h(x),y))$
  - Regression error
  - Multi-class classification error

Learning Goal

- Obtain a good hypothesis, s.t.,
  - h_G if $L_D(h_G) \leq \varepsilon$ (true error)

- Value of ε will vary depending on difficulty of the problem
  - Cannot be less than Bayes error
  - Bayes error is the minimum possible error

New Learning Goal

- Good hypothesis (1)
  - h_G if $L_D(h_G) \leq$ Bayes error + ε (Bayes consistent)
  - Not worse than ε wrt lowest possible error

- Good hypothesis (2)
  - h_G if $L_D(h_G) \leq \min_{h \in H} L_D(h) + \varepsilon$ (Consistent)
  - Not worse than ε wrt the best approximator in H
  - Agnostic learning

Agnostic PAC Learnability

Definition 1 (Agnostic PAC Learnability). *A hypothesis* $\mathcal{H}$ *is agnostic PAC learnable if for every* $\epsilon, \delta \in (0,1)$, *there exists a function* $n_{\mathcal{H}}(\epsilon, \delta)$ *and a learning algorithm such that for every distribution* $\mathcal{D}$ *over* $\mathcal{X} \times \mathcal{Y}$, *if the algorithm is run on* $n \geq n_{\mathcal{H}}(\epsilon, \delta)$ *samples drawn i.i.d. from* $\mathcal{D}$, *then the algorithm returns a hypothesis* $\hat{h}$ *with* $R(\hat{h}) \leq \min_{h \in \mathcal{H}} R(h) + \epsilon$, *except with probability* δ.

Claim

- Finite hypothesis classes are agnostic PAC learnable

Representative Sample

DEFINITION 4.1 (ϵ-representative sample) A training set S is called ϵ-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if

$$\forall h \in \mathcal{H}, \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon.$$

ERM is Successful for Representative Samples

LEMMA 4.2 *Assume that a training set* S *is* $\frac{\epsilon}{2}$*-representative (w.r.t. domain* Z*, hypothesis class* $\mathcal{H}$*, loss function* ℓ*, and distribution* $\mathcal{D}$*). Then, any output of* $\mathrm{ERM}_{\mathcal{H}}(S)$*, namely, any* $h_S \in \mathrm{argmin}_{h\in\mathcal{H}} L_S(h)$*, satisfies*

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

Proof For every $h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h_S) \le L_S(h_S) + \tfrac{\epsilon}{2} \le L_S(h) + \tfrac{\epsilon}{2} \le L_{\mathcal{D}}(h) + \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = L_{\mathcal{D}}(h) + \epsilon,$$

where the first and third inequalities are due to the assumption that S is $\frac{\epsilon}{2}$-representative (Definition 4.1) and the second inequality holds since h_S is an ERM predictor. □

How to get a representative sample

- Sufficiently large (iid) samples are ε-representative

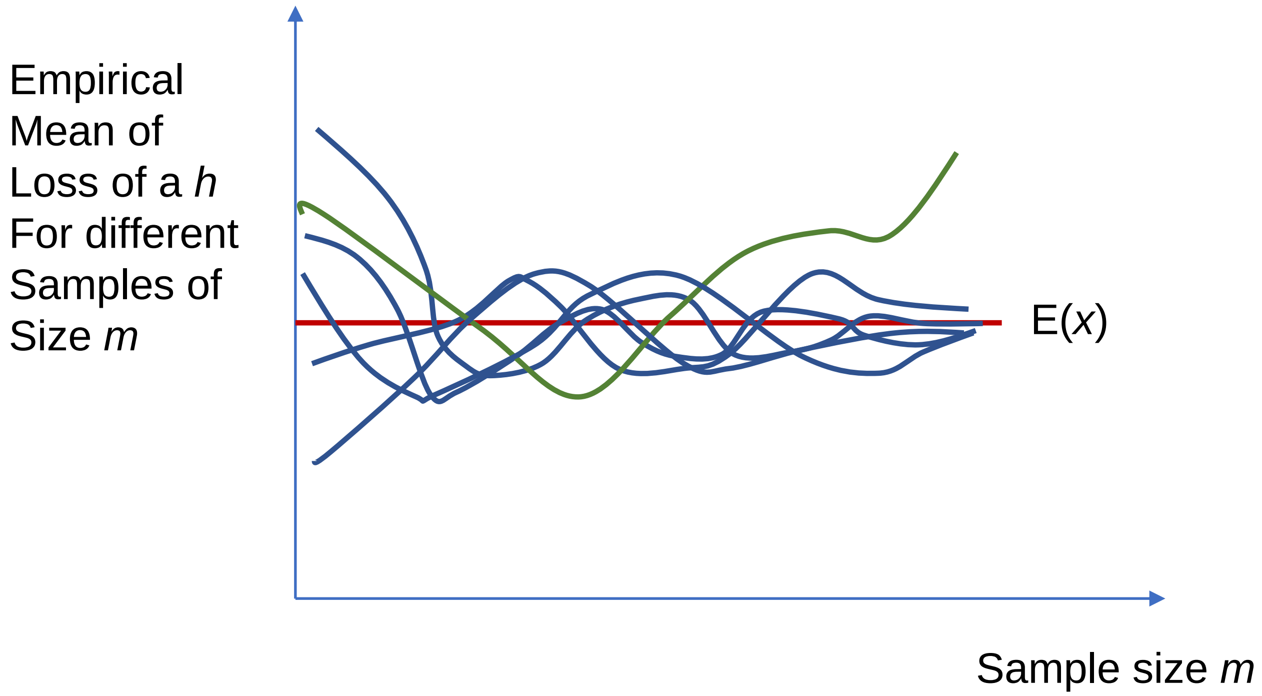
- Empirical risk = $L_{Sm}(h)$
  - empirical mean of random variable $\ell(h(x),y))$ calculated on sample S_m
- True risk = Expectation of random variable $\ell(h(x),y))$
- As sample size grows the empirical risk estimate converges to the true risk

Law of Large Numbers

- As $m \to \infty$ empirical (sample) mean of a random variable converges to its expected value (true mean)

- Here the random variable is $\ell(h(x),y))$ for a particular h

Convergence in Probability



With a high probability empirical mean converges to expectation as sample size grows

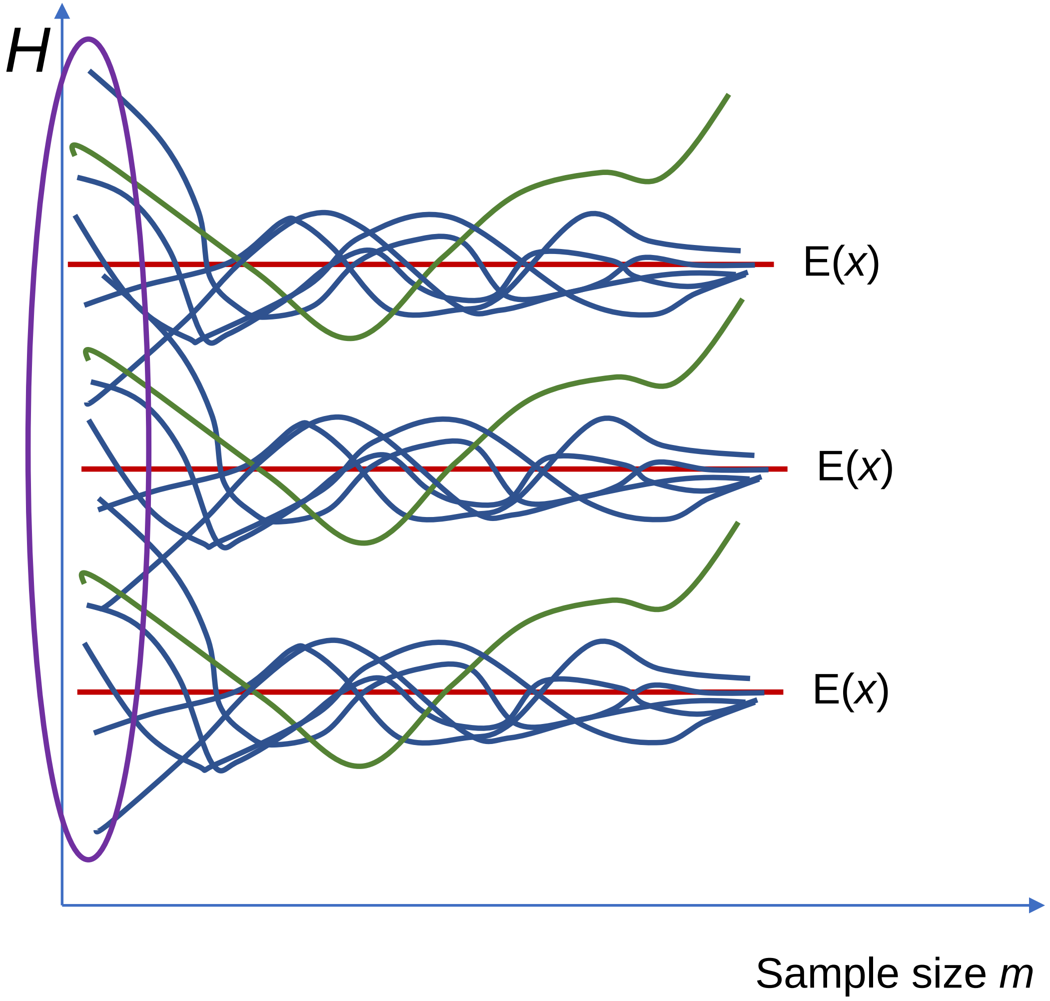
Estimation error is small with
High probability for large samples

Uniform Convergence

Empirical
Mean of
Loss of *h*
For different
Samples of
Size *m*

For EVERY

$h \in H$



Rate of Convergence for Finite Samples

- Concentration of measure inequalities

LEMMA 4.5 (Hoeffding's Inequality) *Let* $\theta_1, \ldots, \theta_m$ *be a sequence of i.i.d. random variables and assume that for all* i, $\mathbb{E}[\theta_i] = \mu$ *and* $\mathbb{P}[a \le \theta_i \le b] = 1$. *Then, for any* $\epsilon > 0$

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \;\le\; 2\exp\left(-2\,m\,\epsilon^2/(b-a)^2\right).$$

Uniform Convergence Property

- If for every $h \in H$ the empirical loss converges to the true loss as sample size goes to infinity the function class H is said to have the property of uniform convergence with respect to distribution D and loss function $l(h(x), y)$.
  - Glivenko-Cantelli class

Definition

DEFINITION 4.3 (Uniform Convergence) We say that a hypothesis class $\mathcal{H}$ has the *uniform convergence property* (w.r.t. a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{\mathrm{UC}} : (0,1)^2 \rightarrow \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{\mathrm{UC}}(\epsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1 - \delta$, S is ϵ-representative.

Agnostic PAC Learnability

COROLLARY 4.4 *If a class $\mathcal{H}$ has the uniform convergence property with a function $m^{UC}_{\mathcal{H}}$ then the class is agnostically PAC learnable with the sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m^{UC}_{\mathcal{H}}(\epsilon/2, \delta)$. Furthermore, in that case, the $\mathrm{ERM}_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.*

Claim

- Finite Hypothesis classes have Uniform Convergence property
- We need to show -

$$\mathcal{D}^m(\{S : \forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon\}) \geq 1 - \delta.$$

Equivalently, we need to show that

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) < \delta.$$

Proof

Writing

$$\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\} = \cup_{h\in\mathcal{H}}\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\},$$

and applying the union bound (Lemma 2.2) we obtain

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{h\in\mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}).$$

Proof (contd.)

$$\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) = \mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \;\leq\; 2\,\exp\left(-2\,m\,\epsilon^2\right). \tag{4.2}$$

Combining this with Equation (4.1) yields

$$\begin{aligned}\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) &\leq \sum_{h\in\mathcal{H}} 2\,\exp\left(-2\,m\,\epsilon^2\right) \\ &= 2\,|\mathcal{H}|\,\exp\left(-2\,m\,\epsilon^2\right).\end{aligned}$$

Proof (Contd.)

Finally, if we choose

$$m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$$

then

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_\mathcal{D}(h)| > \epsilon\}) \leq \delta.$$

COROLLARY 4.6 *Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain, and let $\ell : \mathcal{H} \times Z \to [0, 1]$ be a loss function. Then, $\mathcal{H}$ enjoys the uniform convergence property with sample complexity*

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil .$$

Furthermore, the class is agnostically PAC learnable using the ERM algorithm with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)}{\epsilon^2} \right\rceil .$$