EMPIRICAL RISK MINIMIZATION

Recall from the last lecture:

$$R(f) = \mathbb{E}_{(x,y)\sim P}\left[\ell(f(x), y)\right]$$

for any $f : \mathcal{X} \to \mathcal{Y}$ (measurable). Then, if P was known:

$$f^*(x') \in \underset{y' \in \mathcal{Y}}{\arg\min}\ \mathbb{E}[\ell(y', y) \mid x = x'],\ \forall x' \in \mathcal{X},$$

is a Bayes optimal predictor, i.e., $R(f^*) = R^* = \inf_{\substack{f : \mathcal{X} \to \mathcal{Y} \\ f \text{ is meas.}}} R(f)$

Question: What if P is unknown? Only $(x_i, y_i) \overset{\text{iid}}{\sim} P$, $i = 1, 2, \dots, n$ is provided. $\to$ partial info.

$\to$ Supervised learning

Empirical Risk Minimization (ERM):

Given $S = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, 2, \dots, n\}$, let

$$\hat{R}_S(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i), \qquad f : \mathcal{X} \to \mathcal{Y},\ \text{meas.}$$

$\hat{R}_S(f)$ is the empirical risk of f.

Idea: Use $\hat{f}_{ERM} \in \underset{\substack{f : \mathcal{X} \to \mathcal{Y} \\ f \text{ is meas.}}}{\arg\min}\ \hat{R}_S(f)$.

Hope: The "excess risk" $R(\hat{f}_{ERM}) - R^*$ is small (in expectation, or with high probability, since S and thus $\hat{R}_S$ and $\hat{f}_{ERM}$ are random.

In the following, we show that ERM (empirical risk minimization) is a reasonable idea.

<u>PROPOSITION</u> ($\hat{R}_S(f)$ is an unbiased and consistent estimate of $R(f)$)

Let $f : \mathcal{X} \to \mathcal{Y}$ be a given measurable predictor, and $\ell : \mathcal{Y} \times \mathcal{Y} \to [-B, B]$ for some $B \in \mathbb{R}_+$.

Then,

(i) $$\mathbb{E}[\hat{R}_S(f)] = R(f),$$

(ii) $$\lim_{n\to\infty} \hat{R}_S(f) = R(f) \quad \text{almost surely,}$$

(iii) for any $\delta \in (0,1)$,

$$\mathbb{P}\left(|\hat{R}_S(f) - R(f)| \leq B\sqrt{\frac{2\log(2/\delta)}{n}} \right) \geq 1-\delta.$$

<u>Proof</u> (i) $$\mathbb{E}[\hat{R}(f)] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(f(x_i), y_i)] = \frac{1}{n}\sum_{i=1}^{n} R(f) = R(f),$$

since $(x_i, y_i) \overset{iid}{\sim} P$ for each $i = 1, 2, ---, n$.

(ii) Let $z_i := \ell(f(x_i), y_i)$ for $i = 1, 2, ---, n$. Since $(x_i, y_i) \overset{iid}{\sim} P$, $(z_i)_i$ is an iid sequence, and $\mathbb{E}[z_i] = R(f)$, $\forall i \in [n]$. Also, since $\ell : \mathcal{Y}\times\mathcal{Y} \to [-B, B]$, $|z_i| \leq B$ a.s. for all $i \in [n]$. Thus, by the strong law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} z_i \underset{n\to\infty}{\longrightarrow} \mathbb{E} z_1 = R(f) \quad \text{almost surely.}$$

(iii) By (ii),
$$\mathbb{E}[z_i] = R(f), \ \forall i,$$
$$|z_i| \leq B, \ \text{and } (z_i)_i \text{ is an iid seq.}$$

Thus, by Hoeffding inequality,

$$\mathbb{P}\left(|\hat{R}_S(f) - R(f)| > B\sqrt{\frac{2\log(2/\delta)}{n}} \right)$$
$$= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} z_i - \mathbb{E}[z_1]\right| > B\sqrt{\frac{2\log(2/\delta)}{n}} \right)$$
$$\leq \delta.$$

(Important) Remarks:

1. The performance criterion is $R(f)$, i.e., the population risk. We use $\hat{R}_S$ <u>only</u> as a tool to construct a good predictor $\hat{f}$ to perform well in terms of R.

2. The results of the above proposition is pointwise (for every given f), <u>not</u> uniform that hold simultaneously for all f.

Level End Boss : Overfitting

$\hat{R}_S(f)$ is a consistent and unbiased estimator of $R(f)$, $\forall f$,

and

$$\hat{f}_{ERM} \in \underset{\substack{f : \mathcal{X} \to \mathcal{Y} \\ f \text{ is meas.}}}{\operatorname{argmin}} \hat{R}_S(f).$$

Forget about the computational complexity of finding $\hat{f}_{ERM}$ (for now).

Question : Does $\hat{R}_S(f) = 0$ imply small (or vanishing with n) excess risk $R(f) - R^*$?

The answer is <u>no</u>.

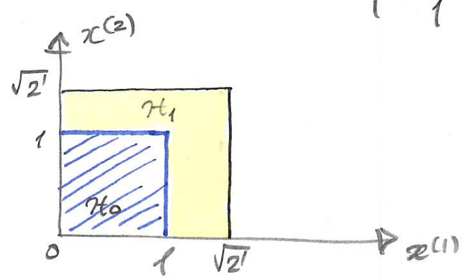
Example : Let $\mathcal{X} = \mathbb{R}^2$, and $\mathcal{Y} = \{0, 1\}$.

The density function of the input is :

$$P_X(x) = \begin{cases} \frac{1}{2}, & x \in [0, \sqrt{2}] \times [0, \sqrt{2}], \\ 0, & \text{otherwise.} \end{cases}$$

Also,

$$y = f^*(x) = \begin{cases} 0, & \text{if } x \in [0,1] \times [0,1], \\ 1, & \text{otherwise.} \end{cases}$$



Consider the following predictor :

$$\hat{f}(x) = \begin{cases} y_i, & \text{if } x = x_i \text{ for some } i = 1, 2, \dots, n \\ 1, & \text{otherwise.} \end{cases}$$

Then, if we consider $\ell(y, y') = \mathbb{1}\{y \neq y'\}$,

$$\hat{R}_S(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\hat{f}(x_i), y_i) = 0.$$

On the other hand,

$$R(\hat{f}) = \int_{\mathcal{H}_0} P_X(x)\, dx = \frac{1}{2}.$$

<u>Perfect fit</u> to the <u>training data</u> implied terrible population risk performance on a <u>test data</u>. $\longrightarrow$ overfitting

Inductive Bias to Avoid Overfitting :

For a given problem, we previously had a humongous hypothesis class :

$$\mathcal{H}^* = \{ f : \mathcal{X} \to \mathcal{Y} \; : \; f \text{ is measurable} \}.$$

This richness resulted in overfitting.

The traditional way to avoid overfitting is to use a restricted hypothesis class. For some well-chosen $\mathcal{H} \subset \mathcal{H}^*$, let

$$\hat{f}_{\mathcal{H}} \in \operatorname*{arg\,min}_{f \in \mathcal{H}} \hat{R}_S(f),$$

and also $R^*_{\mathcal{H}} := \inf_{f \in \mathcal{H}} R(f)$. Then, we have :

$$\boxed{R(\hat{f}_{\mathcal{H}}) - R^* = \underbrace{R(\hat{f}_{\mathcal{H}}) - R^*_{\mathcal{H}}}_{\text{estimation error}} + \underbrace{R^*_{\mathcal{H}} - R^*}_{\text{approximation error}}}$$

excess risk decomposition under inductive bias.

Remarks : There is a tradeoff between the estimation error and the approximation error in the decomposition above :

(i) Large $\mathcal{H}$ $\Rightarrow$ small approximation error
but
large estimation error (larger search)

(ii) Large $n = |S|$ $\Rightarrow$ smaller estimation error
but
no impact on the approximation error.

Some examples :

1. Linear regressors : $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$,

$$\mathcal{H} = \{ x \mapsto \theta^T x \; : \; \theta \in \mathbb{R}^d, \; \|\theta\|_2 \leq \rho \}$$

2. Linear classifiers : $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$,

$$\mathcal{H} = \{ x \mapsto \operatorname{sgn}(\theta^T x - b) \; : \; (\theta, b) \in \mathbb{R}^{d+1} \}$$

③ Quantized perceptrons : $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, +1\}$.

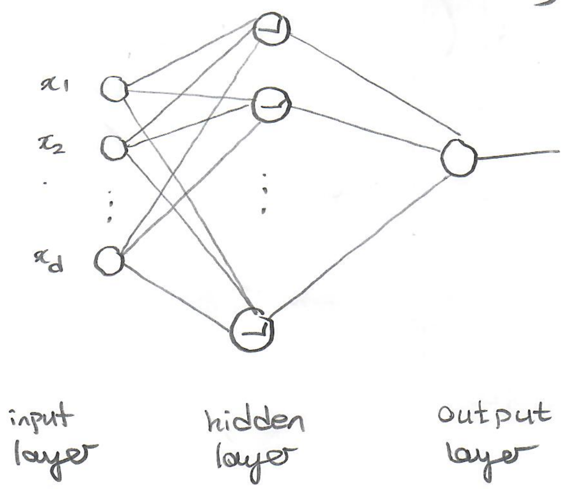
$$\mathcal{H} = \{ x \mapsto \mathrm{sgn}(\omega^T x - b) \ : \ \omega_1, \omega_2, \dots, \omega_d, b \in \mathbb{F}_k \text{ for some } k \in \mathbb{N} \}$$

where $\mathbb{F}_k = \{0,1\}^k$, k

④ Shallow neural networks :

$$\mathcal{H} = \{ x \mapsto \sum_{i=1}^{m} c_i \, \sigma(\omega_i^T x - b_i) \ : \ c, b \in \mathbb{R}^m, \ \omega \in \mathbb{R}^{m \times d} \}$$

$$\omega = \begin{bmatrix} - \omega_1^T - \\ \vdots \\ - \omega_m^T - \end{bmatrix}$$



Capacity Control and Explicit Regularization :

As we mentioned, the traditional way to address overfitting is to consider a restricted hypothesis class $\mathcal{H} \subsetneq \mathcal{H}^*$.

Parameterization : Idea is to define the hypothesis class in a parametric way. Let Θ be a parameter set, and

$$\mathcal{H}_\Theta = \{ x \mapsto f_\theta(x) \in \mathcal{Y} \ : \ x \in \mathcal{X}, \ \theta \in \Theta \}$$

is a parametric hypothesis class.

Then,

$$R^*_{\mathcal{H}_\Theta} = \inf_{f \in \mathcal{H}_\Theta} R(f) = \inf_{\theta \in \Theta} R(f_\theta),$$

$$\hat{\theta}_{ERM} \in \operatorname*{argmin}_{\theta \in \Theta} \hat{R}_S(f_\theta).$$

Examples : ① $\mathcal{H}_\Theta = \{ x \mapsto \theta^T x \ : \ \theta \in \Theta \}$ where $\Theta \subset \mathbb{R}^d$.

② $\Theta \subset \mathbb{R}^{m \times d} \times \mathbb{R}^m \times \mathbb{R}^m$, $m \in \mathbb{Z}_+$,

$$\mathcal{H}_\Theta = \{ x \mapsto \sum_{i=1}^{m} c_i \, \sigma(\omega_i^T x - b_i) \ : \ (\omega, c, b) \in \Theta \}.$$

Recall that the rationale was to control the richness of the hypothesis class, $\mathcal{H}$ or $\mathcal{H}_\Theta$.

Question: Can we further control the richness of $\mathcal{H}_\Theta$? "capacity control."

Answer: Yes. Via (explicit) regularization.

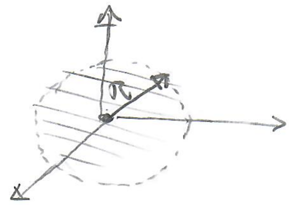
EXAMPLE

Consider
$$\mathcal{H}_\Theta = \{x \mapsto \theta^T x : \theta \in \Theta \subset \mathbb{R}^d\},$$
$$R(f) = \mathbb{E}_{(x,y)\sim P}\left[(\theta^T x - y)^2\right].$$

How to control the richness?

(1)
$$\mathcal{H}_{\Theta,R} = \{x \mapsto \theta^T x : \theta \in \Theta \subset \mathbb{R}^d, \underbrace{\|\theta\|_2 \le R}\},$$
by a design parameter $R > 0$.

Hypotesis class is now a ball of radius $R>0$ around the origin in d-dimensional Euclidean space.



increasing R: (i) decreases the approximation error
$$\inf_{f \in \mathcal{H}_{\Theta,R}} R(f) - R^*$$
since $\mathcal{H}_{\Theta,R} \subset \mathcal{H}_{\Theta,R'}$ for any $R \le R'$.

(ii) increases the optimization and generalization errors, since we have a larger set of candidates.

(2) (Tikhonov regularization)

For a design parameter $\lambda > 0$, consider
$$R^\lambda(f) = R(f) + \lambda \cdot \|\theta\|_2^2, \quad f \in \mathcal{H}_\Theta.$$

In this case, the (regularized) objective is
$$\min_{\theta \in \Theta = \mathbb{R}^d} R^\lambda(f_\theta).$$

As you notice, $\lambda > 0$ implies larger penalty for large $\|\theta\|_2$.

$\Rightarrow$ incentivizes using small $\theta \in \mathbb{R}^d$ $\longrightarrow$ (Norm-based control)

PAC (Probably Approximately Correct) Learnability :

In a given restricted hypothesis class $\mathcal{H} \subset \mathcal{H}^*$, the goal is to come up with $\hat{f} \in \mathcal{H}$ s.t.

$$\mathbb{P}\left(R(\hat{f}) \leq R^*_{\mathcal{H}} + \varepsilon \right) \geq 1 - \delta, \qquad \varepsilon > 0, \ \delta \in (0,1),$$

by using a sufficiently large but finite training set S.

Question : Is this possible for $\mathcal{H} \subset \mathcal{H}^*$?

DEF) (Learning algorithm) For a given learning problem specified by $(\mathbb{X}, \mathbb{Y}, P)$, and a class of hypotheses $\mathcal{H} \subset \mathcal{H}^*$, a learning algorithm is a sequence $A = (A_n)_{n=1}^{\infty}$ of mappings

$$A_n : (\mathbb{X} \times \mathbb{Y})^n \longrightarrow \mathcal{H}.$$

DEF) (PAC - Learnability) Let $\mathcal{H} \subset \mathcal{H}^*$. $\mathcal{H}$ is PAC-learnable if: there is a learning algorithm A such that :

for any $\varepsilon > 0$, $\delta \in (0,1)$, there is an integer $n_0(\varepsilon, \delta)$ s.t.

for any P on $\mathbb{X} \times \mathbb{Y}$, if for any $n \geq n_0$, $S_n \sim P^n$,

then $$\mathbb{P}\left(R(A_n(S_n)) < R^*_{\mathcal{H}} + \varepsilon \right) > 1 - \delta.$$

Important remark : PAC-learnability does not always hold. For some important hypothesis classes, we will prove their PAC-learnability.

As we will see,

- Any finite hypothesis classes,
- Linear regressors with bounded parameter norm :
  $$\mathcal{H} = \{ x \mapsto \theta^T x \ : \|\theta\|_2 \leq R \},$$
- Neural network with large width $m \geq 0$ in a certain operating regime,
- ReLU networks with bounded parameter norm,

will be PAC-learnable.

Sufficiency of Uniform Convergence for PAC-Learnability:

Now, we are given a training set S, and for a hypothesis class $\mathcal{H} \subset \mathcal{H}^*$, an algorithm L_0 returns $\hat{f}$.

Recall :
$$\hat{f}_{\mathcal{H}} \in \underset{f \in \mathcal{H}}{\operatorname{argmin}} \ \hat{R}_S(f),$$

$$R^*_{\mathcal{H}} = \inf_{f \in \mathcal{H}} R(f), \qquad f^*_{\mathcal{H}} \in \underset{f \in \mathcal{H}}{\operatorname{argmin}} \ R(f)$$

Then, we have the following error decomposition:

$$R(\hat{f}) - R^*_{\mathcal{H}} = \underbrace{R(\hat{f}) - \hat{R}(\hat{f})}_{(1)} + \underbrace{\hat{R}_S(\hat{f}) - \hat{R}_S(\hat{f}_{\mathcal{H}})}_{(2)} + \underbrace{\hat{R}_S(\hat{f}_{\mathcal{H}}) - \hat{R}_S(f^*_{\mathcal{H}})}_{(3)} + \underbrace{\hat{R}_S(f^*_{\mathcal{H}}) - R^*_{\mathcal{H}}}_{(4)}$$

Let's examine these error terms :

(3) : Note that $\hat{f}_{\mathcal{H}} \in \underset{f \in \mathcal{H}}{\operatorname{argmin}} \ \hat{R}_S(f)$, and $f^*_{\mathcal{H}} \in \mathcal{H}$. Thus, this term is <u>non-positive.</u>

(4) : Recall from Lecture :

$$Z_i = \ell(f^*_{\mathcal{H}}(x_i), y_i), \quad i = 1, 2, \dots, n.$$

since $(x_i, y_i)_{i=1}^n$ are iid, $(Z_i)_i$ are iid, also $\mathbb{E}[Z_i] = R^*_{\mathcal{H}}$. Thus,

$$|\hat{R}_S(f^*_{\mathcal{H}}) - R^*_{\mathcal{H}}| \le B\sqrt{\frac{2\log(2/\delta)}{n}} \quad \text{w.p.} \ge 1-\delta.$$

(2) : Finding $\hat{f}_{\mathcal{H}}$ may be an NP-hard problem. This term accounts for the optimization error in finding $\hat{f}_{\mathcal{H}}$.

(1) : Same as (4) ? <u>Not quite.</u>

$$\tilde{Z}_i = \ell(\hat{f}(x_i), y_i), \quad i = 1, 2, \dots, n.$$

Note that $\hat{f}$ is $\sigma(S)$-measurable, thus $\tilde{Z}_i$ and $\tilde{Z}_j$ are correlated for $i \neq j$. Furthermore,

$$\mathbb{E}[\ell(\hat{f}(x_i), y_i)] \neq R(\hat{f}) \quad \text{since } \hat{f} \text{ is}$$

$\sigma(S)$-measurable.

Uniform convergence over $\mathcal{H}$:

$$\underbrace{R(\hat{f}) - R^*_{\mathcal{H}}}_{\text{excess risk}} \leq R(\hat{f}) - \hat{R}_S(\hat{f}) + \hat{R}_S(\hat{f}) - \hat{R}_S(\hat{f}_{\mathcal{H}}) + \hat{R}_S(f^*_{\mathcal{H}}) - R(f^*_{\mathcal{H}})$$

$$\leq |R(\hat{f}) - \hat{R}_S(\hat{f})| + |\hat{R}_S(f^*_{\mathcal{H}}) - R(f^*_{\mathcal{H}})| + \hat{R}_S(\hat{f}) - \hat{R}_S(\hat{f}_{\mathcal{H}})$$

$$\leq 2 \cdot \sup_{f \in \mathcal{H}} |R(f) - \hat{R}_S(f)| + \underbrace{\hat{R}_S(\hat{f}) - \hat{R}_S(\hat{f}_{\mathcal{H}})}_{\text{optimization error}}.$$

THEOREM | (Finite $\mathcal{H}$)

Suppose that $|\mathcal{H}| < \infty$. Then, if $\ell : \mathcal{Y} \times \mathcal{Y} \to [-B, B]$,

$$\mathbb{P}\left(\sup_{f \in \mathcal{H}} |R(f) - \hat{R}_S(f)| \leq B \sqrt{\frac{2 \log(2|\mathcal{H}|/\delta)}{n}} \right) \geq 1 - \delta.$$

Thus,

$$\hat{R}(\hat{f}_{\mathcal{H}}) - R^*_{\mathcal{H}} \leq \frac{2B \sqrt{2 \log(2 \cdot |\mathcal{H}| / \delta)}}{\sqrt{n}} \quad \text{w.p.} \geq 1 - \delta,$$

hence any finite $\mathcal{H}$ is PAC-learnable since

$$n \geq \frac{8 B^2 \log(2 \cdot |\mathcal{H}| / \delta)}{\varepsilon^2}$$

implies $\hat{R}(\hat{f}_{\mathcal{H}}) \leq R^*_{\mathcal{H}} + \varepsilon$ w.p. $\geq 1 - \delta$.