INF8953CE - Fall 2020

Machine Learning

- Sarath Chandar

4. Empirical Risk Minimization.

4. Empirical Risk Minimization

Let $x \in \mathbb{R}^p$, $y \in \mathbb{R}$

$h : x \longrightarrow y$ [hypothesis]

When we use model based algorithms, model defines a function 'f' parameterized by 'w'. For every possible parameter configuration w^+, $f(w^+)$ is a hypothesis. Finding right set of parameters for the model is equivalent to finding the right hypothesis from the given set of hypothesis supported by f.

$$EPE(h) = E[L(y, h(x))]$$

$$= \iint L(y, h(x)) P(x, y) \, dx \, dy$$

This is also known as the risk associated with the hypothesis h: $R(h)$

Goal: Find a hypothesis h^* among the class of functions for which the risk $R(h)$ is minimal.

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; R(h)$$

In general, $R(h)$ cannot be computed because the distribution $P(x, y)$ is unknown to the learning algorithm.

However, we can compute an approximation for this risk by using the training set.

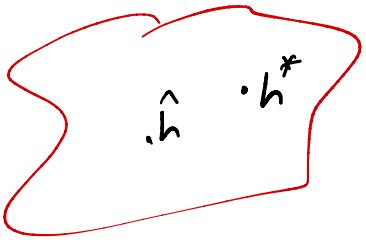
$$R_{emp}(h) = \frac{1}{N} \sum_{i=1}^{N} L(y^{(i)}, h(x^{(i)}))$$

$R_{emp}(h)$: empirical risk.

ERM: find $\hat{h}$ which minimizes the empirical risk.

$$\hat{h} = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; R_{emp}(h)$$

Note 1: There is no guarentee that $h^* = \hat{h}$.



Note 2: The inductive bias of the algorithm restricts the hypothesis class.

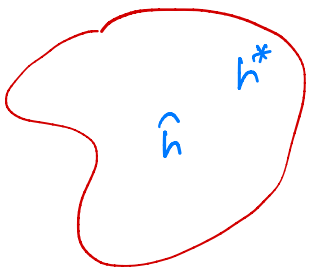
ex: linear models restrict the hypothesis class to linear functions only.

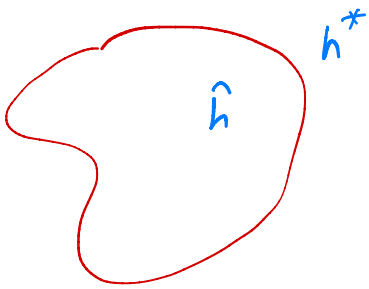
Original hypothesis class: $\mathcal{H}$

restricted hypothesis class: $\mathcal{H}' \subseteq \mathcal{H}$ (subset)

The restricted hypothesis class may or may not contain h^*.

Case 1: $h^* \in \mathcal{H}'$ Case 2: $h^* \notin \mathcal{H}'$





Bias - Variance decomposition:-

Consider regression problem:

$$L(y, h(x)) = (h(x) - y)^2$$

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \; E[L(y, h(x))]$$

$$\boxed{h^* = E[y|x]}$$ regression function.

$$\{h(x) - y\}^2 = \{h(x) - E[y|x] + E[y|x] - y\}^2$$

$$= (h(x) - E[y|x])^2 + (E[y|x] - y)^2$$

$$+ 2\,(h(x) - E[y|x])(E[y|x] - y)$$

$$E[L(y, h(x))] = E[\{h(x) - y\}^2]$$

$$= \underbrace{E[(h(x) - E[y|x])^2]}_{①} + \underbrace{E[(E[y|x] - y)^2]}_{②}$$

$$+ 2\,\underbrace{E[(h(x) - E[y|x])(E[y|x] - y)]}_{③}$$

First, we will show that ③ = 0.

$$③ = E_x E_{y|x}\left(h(x)E[y|x] - h(x)y - E[y|x]^2 + yE[y|x]\right)$$

$$= E_x\left(h(x)E[y|x] - h(x)E[y|x] - E[y|x]^2 + E[y|x]^2\right)$$

$$= E_x[0] = 0.$$

$$E[L(y, h(x))] = \underbrace{E\left[(h(x) - E[y|x])^2\right]}_{①} + \underbrace{E\left[(E[y|x] - y)^2\right]}_{②}$$

① $E\left[(h(x) - E[y|x])^2\right]$

$$= \int_x \int_{y|x} (h(x) - E[y|x])^2 \, p(y|x)\, dy \, p(x)\, dx$$

↳ Not a function of y.

$$= \int_x (h(x) - E[y|x])^2 \, p(x)\, dx$$

② $E\left[(E[y|x] - y)^2\right]$

$$= \int_x \int_{y|x} (y - E[y|x])^2 \, p(y|x)\, dy \, p(x)\, dx$$

$$= \int_x Var(y|x)\, p(x)\, dx.$$

$$E[L(y, h(x))] = E\left[\{h(x) - y\}^2\right]$$

$$= E_x\left[(h(x) - E[y|x])^2\right]$$

$$+ E_x\left[Var(y|x)\right]$$

Note 1: Second term is the variance of the distribution y, averaged over x. It represents the intrinsic variable of the target data and can be regarded as noise.

Second term is independent of $h(x)$

↳ Irreducible minimum value of the loss fn.

if $h(x) = E[y|x]$, then (1) $= 0$

Note 2: First term — is a function of $h(x)$.

This is minimum only when $h(x) = E[y|x]$.

This 0 happens only when we know the true fuction. Even Equ(1) is zero then also we have some irreducible error(variance-noise)

Consider : $\left(h(x) - E[y|x]\right)^2$

In practice, we have a dataset $\mathcal{D}$ containing only a finite number N of data points.

For any dataset $\mathcal{D}$, we run our algorithm and find $h(x)$. Call it $h(x; \mathcal{D})$.

$$\left\{ h(x; \mathcal{D}) - E[y|x] \right\}^2$$

$$E_D\left[\{h(x;D) - E[y|x]\}^2\right]$$

$$= E_D\left[\{h(x;D) - E_D[h(x;D)] + E_D[h(x;D)] - E[y|x]\}^2\right]$$

$$= E_D\left[\{h(x;D) - E_D[h(x;D)]\}^2 + \{E_D[h(x;D)] - E[y|x]\}^2 + 2\,(h(x;D) - E_D[h(x;D)])\,(E_D[h(x;D)] - E[y|x])\right]$$

$= 0$ [Exercise]

$$= \{E_D[h(x;D)] - E[y|x]\}^2 + E_D\left[\{h(x;D) - E_D(h(x;D))\}^2\right]$$

$$= \text{bias}^2 + \text{variance}$$

bias - extent to which the average prediction over all datasets vary around the average.

Variance - extent to which solutions for individual datasets vary around their average.

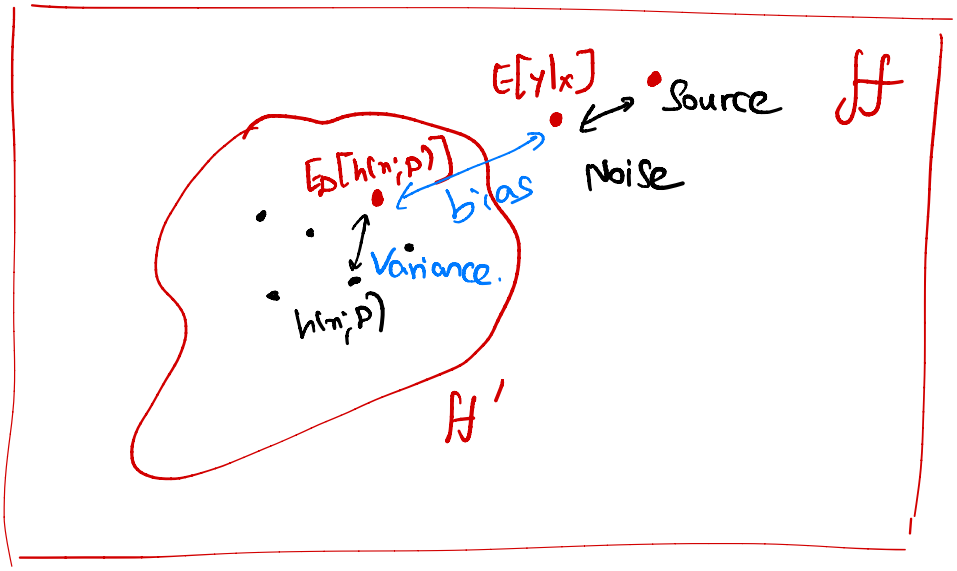
↳ extent to which $h(x;D)$ is sensitive to particular choice of dataset.

Expected loss = $(\text{bias})^2$ + Variance + noise.

$$(\text{bias})^2 = \int \{E_D[h(x;D)] - E[y|x]\}^2 p(x)\,dx$$

$$\text{Variance} = \int \{E_D[h(x;D) - E_D[h(x;D)]]\}^2 p(x)\,dx$$

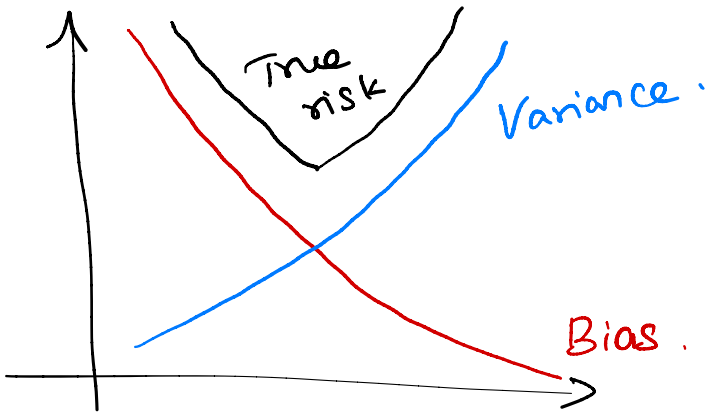
$$\text{noise} = \int \{E[y|x] - y\}^2 p(x,y)\,dx\,dy.$$



Note: There is a tradeoff between bias and Variance.

Very flexible models – low bias and high variance.

rigid models – high bias and low variance.



Occam's Razor :-

"One should not increase, beyond what is necessary, the number of entities required to explain anything".

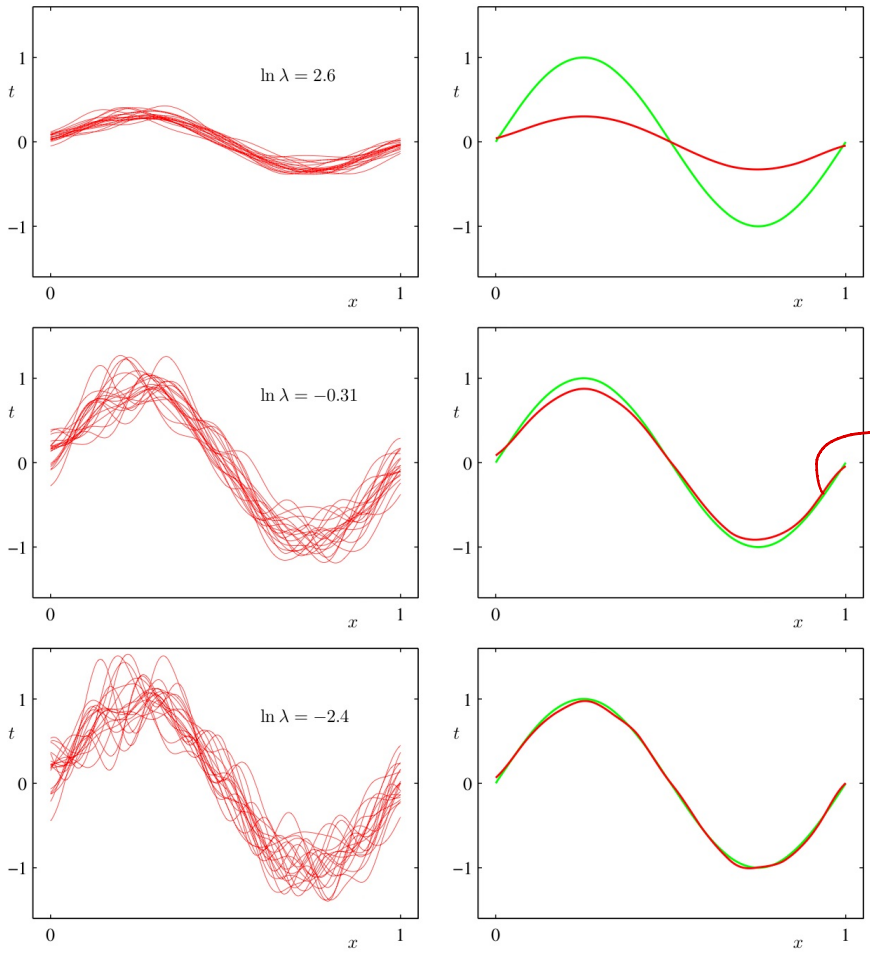
"Seek the simplest explanation".

Regression example :-

100 datasets each containing $N = 25$ data points from $\sin(2\pi x)$ curve.

L = # of datasets = 100

Model : 24 Gaussian basis functions.



λ is large.

variance → low

bias → high

red line
↳ average fit.

λ is small.

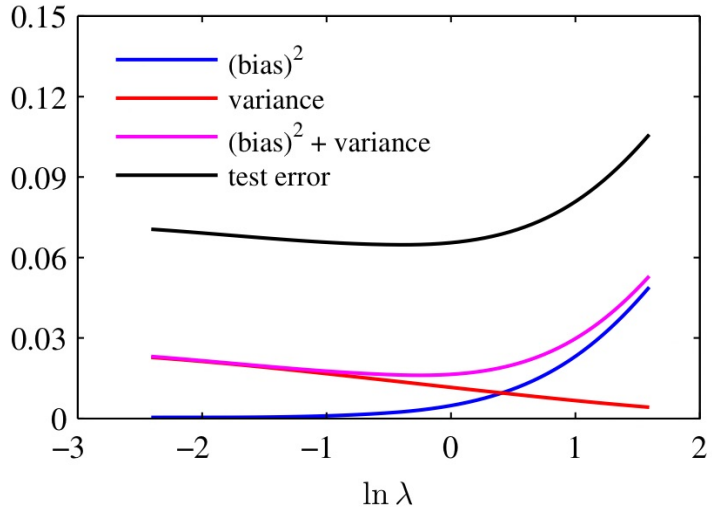
variance → high

bias → low.

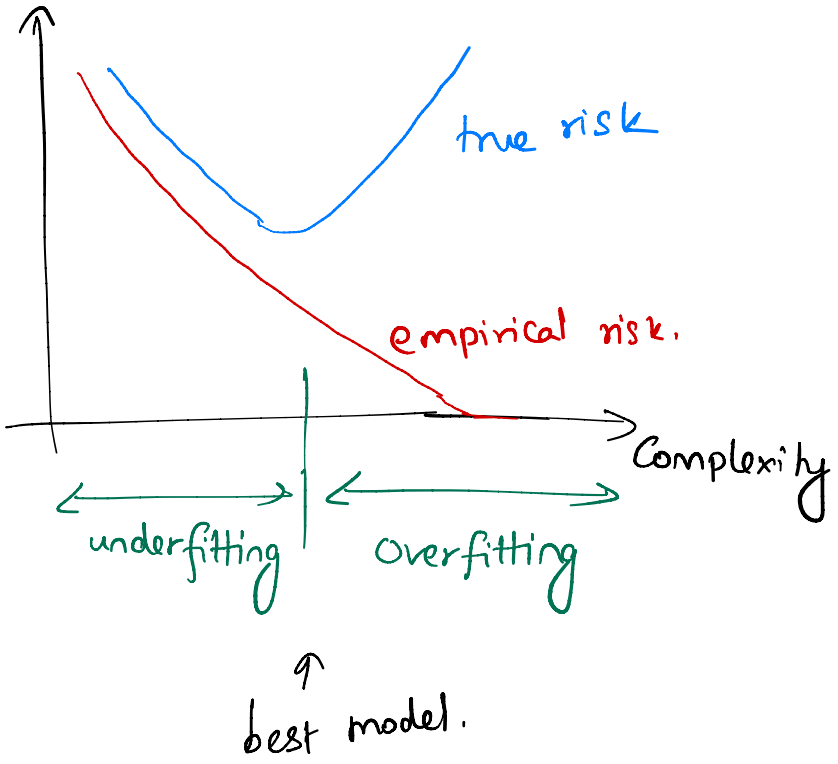
Average prediction : $\bar{h}(x) = \frac{1}{L} \sum_{l=1}^{L} h^{(l)}(x)$

$$(\text{bias})^2 = \frac{1}{N} \sum_{n=1}^{N} \left\{ \bar{h}(x^{(n)}) - E[y^{(n)} | x^{(n)}] \right\}^2$$

$$\text{Variance} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{L} \sum_{l=1}^{L} \left\{ h^{(l)}(x^{(n)}) - \bar{h}(x^{(n)}) \right\}^2$$



Relation between true risk, empirical risk, overfitting :-



You should know!

1. Hypothesis
2. True risk
3. Empirical risk.
4. Bias- variance decomposition.
5. Irreducible error.
6. Occam's razor.

7. underfitting / overfitting / true risk / empirical risk / Complexity of the model – Connections.
