BIAS - VARIANCE TRADE-OFF I

1) Hypothesis space: $\mathcal{H}$

- $\mathcal{H}$ is a compact subset of $C(X)$, equipped with the standard norm

$$\|f\|_{C(X)} := \|f\|_{\infty} := \sup_{\vec{x} \in X} |f(\vec{x})|$$

2) Empirical target function:

- h^* is not computable
- h^* may not reside in $\mathcal{H}$

$\Rightarrow$ Thus the best we hope for is to find a target function $\hat{h}$ in $\mathcal{H}$ that is closest to h^*

$$\hat{h} := \underset{\mathcal{H}}{\operatorname{argmin}} \int_X \left(f(x) - h^*(x)\right)^2 d\pi(\vec{x})$$

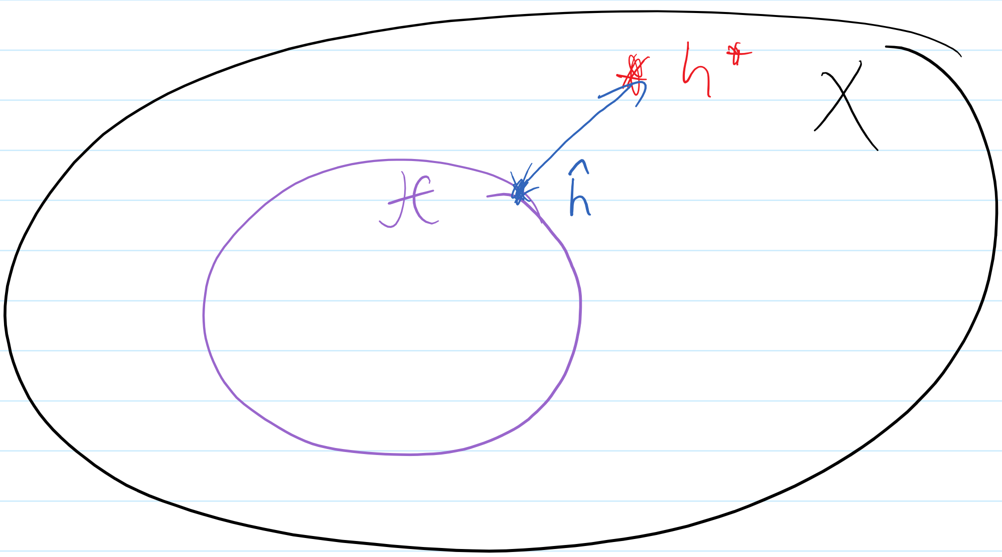
Ex 17.1. $\hat{h}$ is also closest to y, that is

$\hat{h}$ is also a minimizer of

$$\min_{\mathcal{H}} \int_Z \left(f(\vec{x}) - y\right)^2 d\pi(\vec{x}, y)$$

$$\Downarrow$$

$$\hat{h} := \underset{\mathcal{H}}{\operatorname{argmin}} \ \mathcal{R}(h)$$



- Again we know $\pi(\vec{x}, y)$ only through the training set $\Rightarrow$ $\hat{h}$ is NOT computable. Thus we resort to empirical target function:

$$\boxed{\hat{h}_N := \underset{f \in \mathcal{H}}{\operatorname{argmin}} \ \mathcal{R}_N(f) = \frac{1}{N} \sum_{i=1}^{N} \left(f(\vec{x}^i) - y^i\right)^2}$$

EXISTENCE OF $\hat{h}$, $\hat{h}_N$

Assumption: (M- Boundedness of the misfit on $\mathcal{H}$)

for any $h \in \mathcal{H}$ and a.s. (a.e) in X

there holds:
$$|h(\vec{x}) - y| \leq M$$

Prop 17.1: Suppose $\mathcal{H}$ is M-bounded. Then $\mathcal{R}, \mathcal{R}_N : \mathcal{H} \to \mathbb{R}$ are Lipschitz continuous, i.e.

$$|\mathcal{R}(h_1) - \mathcal{R}(h_2)| \leq c \, \|h_1 - h_2\|_\infty$$
$$|\mathcal{R}_N(h_1) - \mathcal{R}_N(h_2)| \leq c \, \|h_1 - h_2\|_\infty$$

Proof:
$$|\mathcal{R}(h_1) - \mathcal{R}(h_2)| = \int_Z \left[(h_1 - y)^2 - (h_2 - y)^2\right] d\pi$$

$$\|$$

$$\left| \int_Z (h_1 + h_2 - 2y)(h_1 - h_2) \, d\pi \right|$$

$\wedge\!\!/\!\!/$ M- boundedness

$$2M \, \|h_1 - h_2\|_\infty$$

Cor. 17.1: Suppose $\mathcal{H}$ is M-bounded, then $\hat{h}$ and $\hat{h}_N$ exist.

Proof: direct consequence of Weirstrass theorem and Prop 17.1

BIAS - VARIANCE TRADE-OFF

- We are interested in the the actual risk $R(\hat{h}_N)$. Let us start with

$$R(\hat{h}_N) = \underbrace{R(\hat{h}_N) - R(\hat{h})}_{S(\hat{h}_N)} + \underbrace{R(\hat{h})}_{\overset{!!}{=}\, B(\hat{h})}$$

Goal: is to bound $S(\hat{h}_N)$ and $B(\hat{h})$

- Let us first consider

$$S(\hat{h}_N) = R_N(\hat{h}_N) - R_N(\hat{h}) + R(\hat{h}_N) - R_N(\hat{h}_N) + R_N(\hat{h}) - R(\hat{h})$$

$$\leq \left| R(\hat{h}_N) - R_N(\hat{h}_N) \right| + \left| R_N(\hat{h}) - R(\hat{h}) \right|$$

↑ because $\hat{h}_N$ is a minimizer of R_N

$\Rightarrow$ $S(\hat{h}_N)$ is bounded by sampling errors $\Rightarrow$ hence the name. $S(\hat{h}_N)$ is also known as the variances.

- Next let us look at

$$B(\hat{h}) := R(\hat{h}) \overset{??}{:=} \int_X \left(\hat{h}(\vec{x}) - h^*(\vec{x})\right)^2 d\pi(\vec{x}) + \sigma_\pi^2$$

$$\Downarrow$$

The bias depends only the approximation capability of the hypothesis space $\mathcal{H}$ [closer h^* to $\mathcal{H}$ the better the error].

$$\Downarrow$$

the task of approximation theory (NOT a focus of this class)

* Summary: we focus on estimating the sampling/variance error in this class

* Bias - Variance trade-off

* Bias - Variance trade-off

- For fixed $\mathcal{H}$, the sampling error decreases as we increase the sample size N.

- For a fixed sample size N, enriching the hypothesis space $\mathcal{H}$ reduces the bias

$\Rightarrow$ the popular trade-off is to enlarge the hypothesis space as the sample size increases.