Lecture 4
Bias-Variance Tradeoff
&
Intro to Probabilistic Inference

How to quantify expected *out-of-sample* error?

$$y = f(\boldsymbol{x}) + \epsilon \quad \text{with} \quad \epsilon \in \mathcal{N}(0, \sigma_\epsilon^2)$$

$$\ell(\boldsymbol{y}, h(\boldsymbol{X}; \boldsymbol{\theta})) = \sum_i \left(y_i - h\left(\boldsymbol{x}_i; \boldsymbol{\theta}\right)\right)^2$$

Given a fixed training set D, the empirically optimal hypothesis is parametrised by

$$\hat{\boldsymbol{\theta}}_D = \arg\min_{\boldsymbol{\theta}} \ell(\boldsymbol{y}, h(\boldsymbol{X}; \boldsymbol{\theta}))$$

On average, how well does our empirically optimal hypothesis $h_S^* \equiv h\left(\boldsymbol{x}, \hat{\boldsymbol{\theta}}_D\right)$ *make prediction*?

In other words, we want to quantify how well our empirically optimal hypothesis from finite samples predict the value of the unknown god-given function f.

We'll perform a bias-variance decomposition to do quantify this.

Let's calculate the out-of-sample error, averaged over multiple data samples D and noise realisation ϵ

$$\mathbb{E}_{D,\epsilon}\left[\ell\left(\boldsymbol{y}, h\left(\boldsymbol{X};\hat{\boldsymbol{\theta}}_D\right)\right)\right] = \mathbb{E}_{D,\epsilon}\left[\sum_i \left(y_i - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right)^2\right]$$

Let's calculate the out-of-sample error, averaged over multiple data samples D and noise realisation ϵ

$$\mathbb{E}_{D,\epsilon}\left[\ell\left(\boldsymbol{y}, h\left(\boldsymbol{X};\hat{\boldsymbol{\theta}}_D\right)\right)\right] = \mathbb{E}_{D,\epsilon}\left[\sum_i \left(y_i - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right)^2\right]$$

$$= \mathbb{E}_{D,\epsilon}\left[\sum_i \left(y_i - f\left(\boldsymbol{x}_i\right) + f\left(\boldsymbol{x}_i\right) - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right)^2\right]$$

Let's calculate the out-of-sample error, averaged over multiple data samples D and noise realisation ϵ

$$\mathbb{E}_{D,\epsilon}\left[\ell\left(\boldsymbol{y}, h\left(\boldsymbol{X};\hat{\boldsymbol{\theta}}_D\right)\right)\right] = \mathbb{E}_{D,\epsilon}\left[\sum_i \left(y_i - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right)^2\right]$$

$$= \mathbb{E}_{D,\epsilon}\left[\sum_i \left(y_i - f\left(\boldsymbol{x}_i\right) + f\left(\boldsymbol{x}_i\right) - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right)^2\right]$$

$$= \sum_i \mathbb{E}_{\epsilon}\left[\left(y_i - f\left(\boldsymbol{x}_i\right)\right)^2\right] + \mathbb{E}_{D,\epsilon}\left[\left(f\left(\boldsymbol{x}_i\right) - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right)^2\right]$$

$$+2\mathbb{E}_{\epsilon}\left[y_i - f\left(\boldsymbol{x}_i\right)\right]\mathbb{E}_{D}\left[f\left(\boldsymbol{x}_i\right) - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]$$

Let's calculate the out-of-sample error, averaged over multiple data samples D and noise realisation ϵ

$$
\begin{aligned}
\mathbb{E}_{D,\epsilon}\left[\ell\left(\boldsymbol{y}, h\left(\boldsymbol{X};\hat{\boldsymbol{\theta}}_D\right)\right)\right] &= \mathbb{E}_{D,\epsilon}\left[\sum_i \left(y_i - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right)^2\right] \\
&= \mathbb{E}_{D,\epsilon}\left[\sum_i \left(y_i - f\left(\boldsymbol{x}_i\right) + f\left(\boldsymbol{x}_i\right) - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right)^2\right] \\
&= \sum_i \mathbb{E}_{\epsilon}\left[\left(y_i - f\left(\boldsymbol{x}_i\right)\right)^2\right] + \mathbb{E}_{D,\epsilon}\left[\left(f\left(\boldsymbol{x}_i\right) - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right)^2\right] \\
&\quad + 2\mathbb{E}_{\epsilon}\left[y_i - f\left(\boldsymbol{x}_i\right)\right]\mathbb{E}_{D}\left[f\left(\boldsymbol{x}_i\right) - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right] \\
&= \sum_i \sigma_\epsilon^2 + \sum_i \mathbb{E}_{D}\left[\left(f\left(\boldsymbol{x}_i\right) - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right)^2\right]
\end{aligned}
$$

measurement error (noise) *finite-size sampling error (noise independent)*

Let's look at the last term

finite-size sampling error

$$\sum_i \mathbb{E}_D\left[\left(f(\boldsymbol{x}_i) - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right)^2\right] = \sum_i \mathbb{E}_D\left[\left\{f(\boldsymbol{x}_i) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]} + \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]} - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right\}^2\right]$$

Let’s look at the last term

finite-size sampling error

$$
\begin{aligned}
\sum_i \mathbb{E}_D\left[\left(f(\boldsymbol{x}_i) - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right)^2\right] &= \sum_i \mathbb{E}_D\left[\left\{f(\boldsymbol{x}_i) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]} + \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]} - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right\}^2\right] \\
&= \sum_i \mathbb{E}_D\left[\left\{f(\boldsymbol{x}_i) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right\}^2\right] + \mathbb{E}_D\left[\left\{h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right\}^2\right] \\
&\quad + 2\mathbb{E}_D\left[\left\{f(\boldsymbol{x}_i) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right\}\left\{h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right\}\right]
\end{aligned}
$$

Let's look at the last term

finite-size sampling error

$$\sum_i \mathbb{E}_D\left[\left(f(\boldsymbol{x}_i) - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right)^2\right] = \sum_i \mathbb{E}_D\left[\left\{f(\boldsymbol{x}_i) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]} + \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]} - h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right\}^2\right]$$

$$= \sum_i \mathbb{E}_D\left[\left\{f(\boldsymbol{x}_i) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right\}^2\right] + \mathbb{E}_D\left[\left\{h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right\}^2\right]$$

$$+2\mathbb{E}_D\left[\left\{f(\boldsymbol{x}_i) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right\}\left\{h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right\}\right]$$

$$= \sum_i \left(f(\boldsymbol{x}_i) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right)^2 + \sum_i \mathbb{E}_D\left[\left\{h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right\}^2\right]$$

Bias^2 Variance

How to quantify expected *out-of-sample* error? Putting them altogether

$$\mathbb{E}_{D,\epsilon}\left[\ell\left(\boldsymbol{y}, h\left(\boldsymbol{X};\hat{\boldsymbol{\theta}}_D\right)\right)\right] = \sum_i \left(f\left(\boldsymbol{x}_i\right) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right)^2$$

Bias²

$+$

$$+\sum_i \mathbb{E}_D\left[\left\{h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right\}^2\right]$$

Variance

$+$

$$+\sum_i \sigma_\epsilon^2$$

Measurement error
(Noise)

How to quantify expected *out-of-sample* error? Putting them altogether

$$\mathbb{E}_{D,\epsilon}\left[\ell\left(\boldsymbol{y}, h\left(\boldsymbol{X}; \hat{\boldsymbol{\theta}}_D\right)\right)\right] = \sum_i \left(f\left(\boldsymbol{x}_i\right) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_D\right)\right]}\right)^2$$

Bias^2

+

$$+ \sum_i \mathbb{E}_D\left[\left\{h\left(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_D\right) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_D\right)\right]}\right\}^2\right]$$

Variance

sample size dependent!

+

$$+ \sum_i \sigma_\epsilon^2$$

Measurement error
(Noise)

How to quantify expected *out-of-sample* error? Putting them altogether

$$\mathbb{E}_{D,\epsilon}\left[\ell\left(\boldsymbol{y}, h\left(\boldsymbol{X};\hat{\boldsymbol{\theta}}_D\right)\right)\right] = \sum_i \left(f\left(\boldsymbol{x}_i\right) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right)^2$$

Bias^2

+

$$+\sum_i \mathbb{E}_D\left[\left\{h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right) - \underline{\mathbb{E}_D\left[h\left(\boldsymbol{x}_i;\hat{\boldsymbol{\theta}}_D\right)\right]}\right\}^2\right]$$

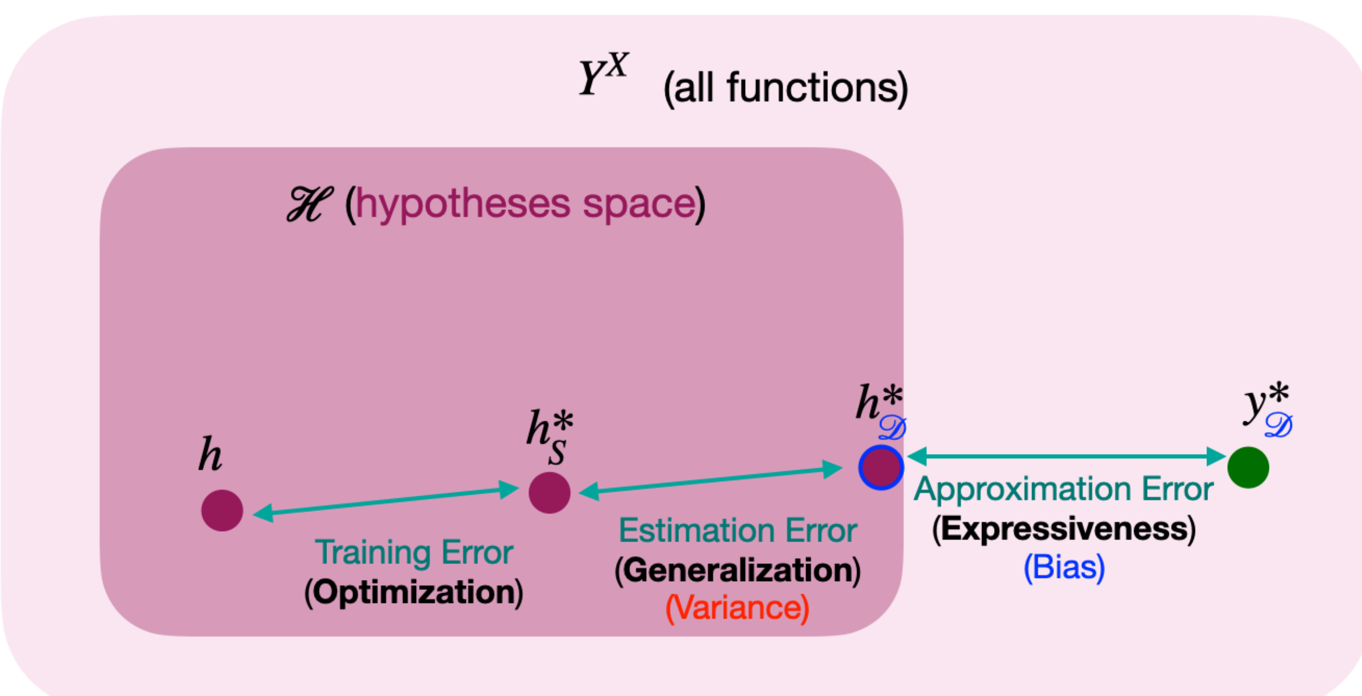
Variance

sample size dependent!

+

$$+\sum_i \sigma_\epsilon^2$$

Measurement error
(Noise)



This picture is valid if the consistency condition holds

$$\underline{\mathbb{E}_D\left[h\left(\boldsymbol{x};\hat{\boldsymbol{\theta}}_D\right)\right]} \approx \lim_{m\to\infty} h_S^* = h_{\mathscr{D}}^*.$$

How to quantify expected *out-of-sample* error? Putting them altogether

$$\mathbb{E}_{D,\epsilon}\left[\ell\left(y, h\left(X;\hat{\theta}_D\right)\right)\right] = \sum_i \left(f\left(x_i\right) - \underline{\mathbb{E}_D\left[h\left(x_i;\hat{\theta}_D\right)\right]}\right)^2$$

Bias²

$$+\sum_i \mathbb{E}_D\left[\left\{h\left(x_i;\hat{\theta}_D\right) - \underline{\mathbb{E}_D\left[h\left(x_i;\hat{\theta}_D\right)\right]}\right\}^2\right]$$

+

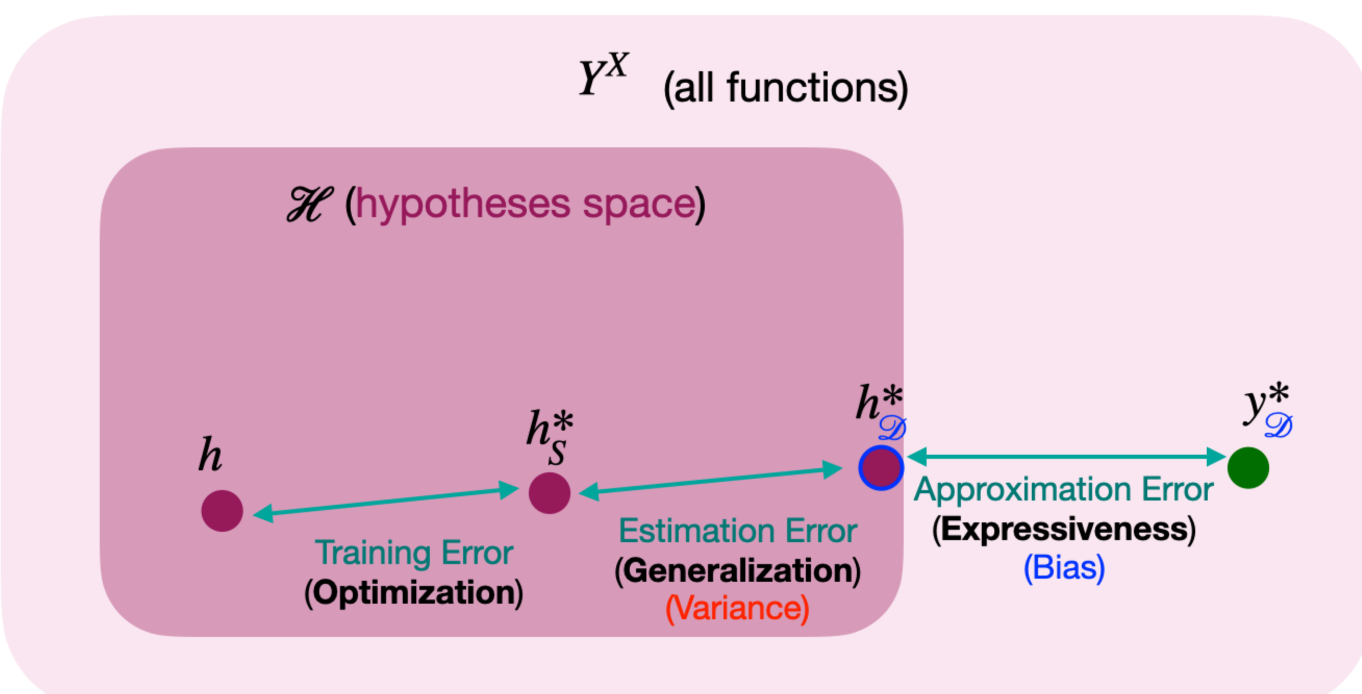
Variance

sample size dependent!

$$+\sum_i \sigma_\epsilon^2$$

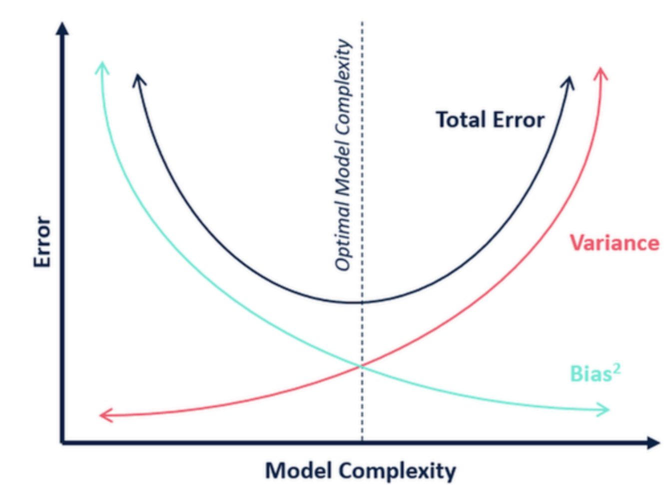
+

Measurement error
(Noise)

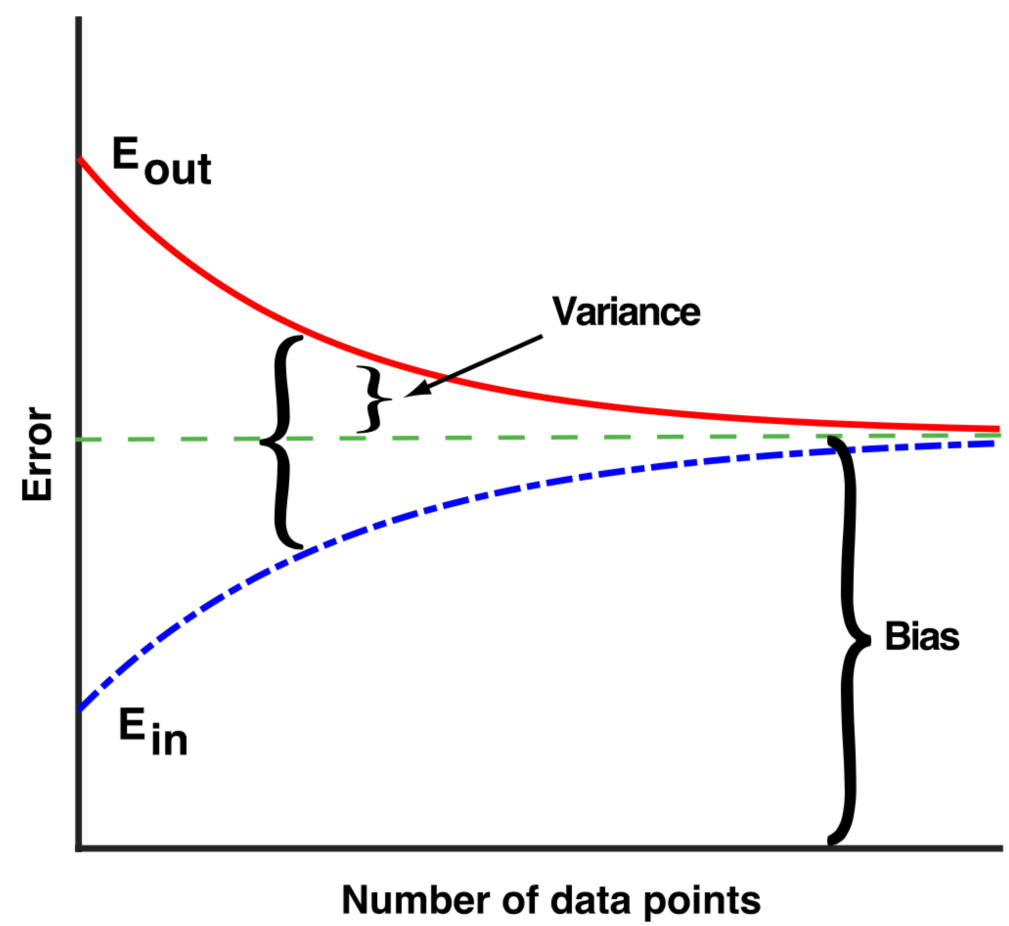


This picture is valid if the consistency condition holds

$$\underline{\mathbb{E}_D\left[h\left(x;\hat{\theta}_D\right)\right]} \approx \lim_{m\to\infty} h_S^* = h_{\mathscr{D}}^*.$$



Note: this graph is not quite right…



Measurement error is troublesome!

It makes out-of-sample non-zero even our hypothesis class the same as the god-given signal!

Can we design an alternative approach to learn both signal and noise?

Yes! Bayesian and Probabilistic Inference

Back to a noisy god-given rule

$$y = f(\boldsymbol{x}) + \epsilon \quad \text{with} \quad \epsilon \in \mathcal{N}(0, \sigma_\epsilon^2)$$

We can write this rule for generating the continuous label as

$$P(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}\left(f(\boldsymbol{x}), \sigma_\epsilon^2\right)$$

This is a *generative model* (probabilistic model that generates data label, given input).

Back to a noisy god-given rule

$$y = f(\boldsymbol{x}) + \epsilon \quad \text{with} \quad \epsilon \in \mathcal{N}(0, \sigma_\epsilon^2)$$

We can write this rule for generating the continuous label as

$$P(y \mid \boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}\left(f(\boldsymbol{x}), \sigma_\epsilon^2\right)$$

This is a *generative model* (probabilistic model that generates data label, given input).

What's the framework to evaluate the hypothesis function?

Frequentist's Interpretation of Probability

$P(A)$ = *long-run relative frequency* with which A occurs in identical repeats of an experiment.
"A" restricted to propositions about *random variables*.

Law of large number relates frequency of events in repeated experiments to theoretical probability.

Frequentist's Interpretation of Probability

$P(A)$ = *long-run relative frequency* with which A occurs in identical repeats of an experiment. "*A*" restricted to propositions about *random variables*.

Law of large number relates frequency of events in repeated experiments to theoretical probability.

Bayesian Interpretation of Probability

$P(A \mid B)$ = a real number measure of *the plausibility of a proposition/hypothesis A*, given (conditional on) *the truth* of the information represented by proposition B.
"A" can be *any logical proposition*, <u>not restricted to</u> propositions about random variables.

Both views follow the same mathematical rules of probability theory…

One useful rule (Bayes' Rule)

$$\boxed{P(A \mid B) = \frac{P(A, B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}}$$

Bayes' rule for evaluating plausibility of a scientific hypothesis, given data.

$$\boxed{P\left(H_i \mid D, I\right) = \frac{P\left(H_i \mid I\right) P\left(D \mid H_i, I\right)}{P(D \mid I)}}$$

H_i = proposition asserting the truth of a hypothesis of interest

I = proposition representing our prior information

D = proposition representing data

$P\left(D \mid H_i, I\right)$ = probability of obtaining data D; if H_i and I are true (Likelihood function)

$P\left(H_i \mid I\right)$ = Prior probability of hypothesis

$P\left(H_i \mid D, I\right)$ = Posterior probability of hypothesis

$P(D \mid I) = \sum_i P\left(H_i \mid I\right) P\left(D \mid H_i, I\right)$ is the normalization (tricky to evaluate)

Bayes' rule for evaluating plausibility of a scientific hypothesis, given data.

$$P\left(H_i \mid D, I\right) = \frac{P\left(H_i \mid I\right) P\left(D \mid H_i, I\right)}{P(D \mid I)}$$

H_i = proposition asserting the truth of a hypothesis of interest

I = proposition representing our prior information

D = proposition representing data

$P\left(D \mid H_i, I\right)$= probability of obtaining data D; if H_i and I are true (Likelihood function)

$P\left(H_i \mid I\right)$ = Prior probability of hypothesis

$P\left(H_i \mid D, I\right)$ = Posterior probability of hypothesis

$P(D \mid I) = \sum_i P\left(H_i \mid I\right) P\left(D \mid H_i, I\right)$ is the normalization (tricky to evaluate)

Bayes' rule allows you to evaluate the probability that the hypothesis is true once new data arrives! (How your belief changes depending on the incoming data)