Foundations of Machine Learning
AI2000 and AI5000

FoML-10
Bias Variance Decomposition

Dr. Konda Reddy Mopuri
Department of AI, IIT Hyderabad
July-Nov 2025

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence
& Learning Lab

So far in FoML

- What is ML and the learning paradigms
- Probability refresher
- MLE, MAP, and fully Bayesian treatment
- Linear Regression with basis functions - and regularization
- Model selection

Breaking down the prediction error of a model

Frequentist interpretation of the model complexity

Expected Loss for Regression

- Regression loss $L(t, y(\mathbf{x})) = [t - y(\underline{x})]^2$

for a given $(\underline{x}, t) \sim P(\underline{x}, t)$

Expected Loss for Regression

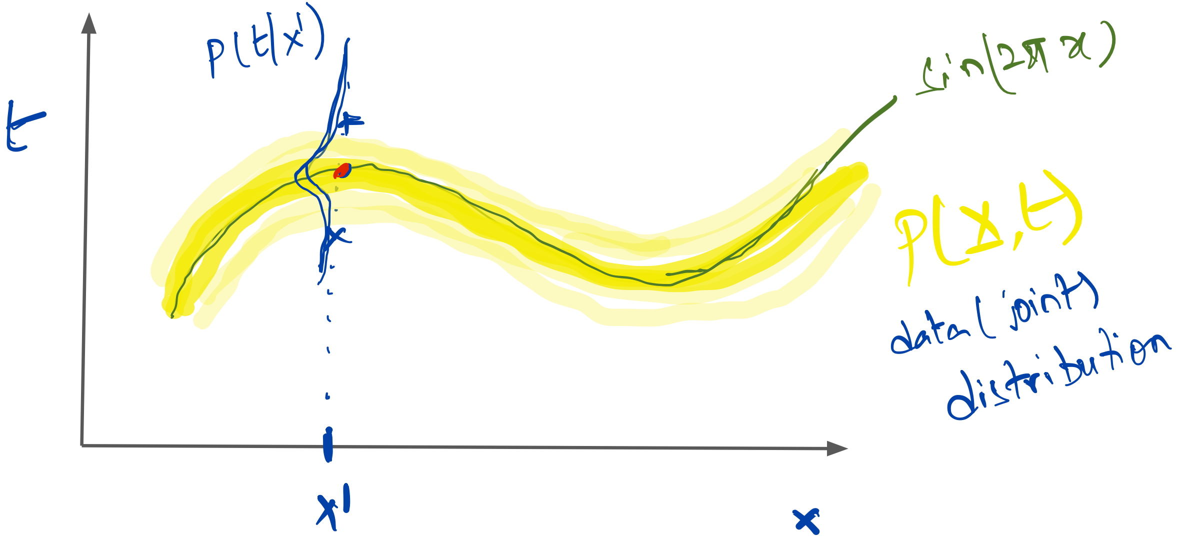
- Regression loss $L(t, y(\mathbf{x})) =$ $[t - y(\underline{x})]^2$

  for a given $(\underline{x}, t) \sim p(\underline{x}, t)$

- If we know the data distribution, we can find the

$$\mathbb{E}[L(t, y((\mathbf{x})))] = \iint (t - y(\underline{x}))^2 \, p(\underline{x}, t) \, d\underline{x} \, dt$$

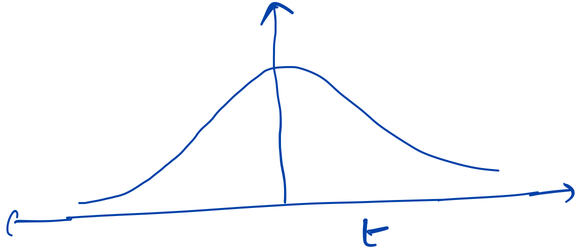
Data and prediction distributions



$$t = \sin 2\pi x + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \beta^{-1})$$

$$p(t|\mathbf{x})$$



Minimizing the Expected loss at given x

$$\mathcal{L} = \int [t - y(\underline{x})]^2 \, p(t|\underline{x}) \, dt$$

$$\frac{\partial L}{\partial y(\underline{x})} = 2 \int [t - y(\underline{x})] \, p(t|\underline{x}) \, dt = 0$$

$$\int t \, p(t|\underline{x}) \, dt = \int y(\underline{x}) \, p(t|\underline{x}) \, dt$$

$$\boxed{E[t|\underline{x}] = y(\underline{x})}$$

Regression function

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Data-driven Intelligence
& Learning Lab

Expected Loss for Regression

$$\mathbb{E}[L] = \int \int (y(\mathbf{x}) - t)^2 p(\mathbf{x}, t) dt d\mathbf{x}$$

$$= \mathbb{E}_{x,t}\left[(y(\underline{x}) - t)^2\right] = \mathbb{E}_{x,t}\left[(y(\underline{x}) - \mathbb{E}[t/\underline{x}] + \mathbb{E}[t/\underline{x}] - t)^2\right]$$

$$= \underbrace{\mathbb{E}_{x,t}\left[(y(\underline{x}) - \mathbb{E}[t/\underline{x}])^2\right]}_{\text{Due to the model}} + \underbrace{\mathbb{E}_{x,t}\left[(\mathbb{E}[t/\underline{x}] - t)^2\right]}_{\text{Due to the intrinsic noise}}$$

Cross terms vanish

$$\mathbb{E}_{x,t}\left[\{y(x) - \mathbb{E}[t/x]\} \cdot \{\mathbb{E}[t/x] - t\}\right]$$

$$= \mathbb{E}_x\left[\mathbb{E}_t\left[\{y(x) - \mathbb{E}(t/x)\} \cdot \underbrace{\{\mathbb{E}[t/x) - t\} \mid x}_{0}\right]\right]$$

Minimizing the expected loss

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t/\mathbf{x}]\}^2 p(\mathbf{x}) d\mathbf{x} + \int \mathrm{var}[t/\mathbf{x}] p(\mathbf{x}) d\mathbf{x}$$

- Optimal solution is unknown $\quad y(\mathbf{x}) = \mathbb{E}[t/\mathbf{x}]$

If we model $\mathbb{E}[t|\mathbf{x}]$ using parameters $\underline{w}$, then from a Bayesian perspective we can express the model's uncertainty via a posterior on $\underline{w}$

But we make point estimate for $\underline{w}$ on a dataset D

Minimizing the expected loss

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t/\mathbf{x}]\}^2 p(\mathbf{x}) d\mathbf{x} + \int \mathrm{var}[t/\mathbf{x}] p(\mathbf{x}) d\mathbf{x}$$

- Optimal solution is unknown $y(\mathbf{x}) = \mathbb{E}[t/\mathbf{x}]$

- We only have finite dataset (but not the distribution)

$D = \{(x_1, t_1) \ldots\ldots (x_N, t_N)\}$

Minimizing the expected loss

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t/\mathbf{x}]\}^2 p(\mathbf{x}) d\mathbf{x} + \int \mathrm{var}[t/\mathbf{x}] p(\mathbf{x}) d\mathbf{x}$$

- Frequentist approach → multiple datasets, multiple models

$D_1 = \{ \cdots \}$ $D_2 = \{ \cdots \}$ $\cdots$ $D_L = \{ \}$

y_1 y_2 y_L

$$\mathbb{E}_D[(y_D(\mathbf{x}) - \mathbb{E}[t/\mathbf{x}])^2]$$

Estimate the performance by averaging the expected loss over different datasets

Minimizing the expected loss

$$\mathbb{E}[\mathbb{E}_D[L]] = \int \mathbb{E}_D[(y_D(\mathbf{x}) - \mathbb{E}[t/\mathbf{x}])^2]p(\mathbf{x})d\mathbf{x} + \int \mathrm{var}[t/\mathbf{x}]p(\mathbf{x})d\mathbf{x}$$

- Bias-Variance decomposition

Minimizing the expected loss

$$\mathbb{E}[\mathbb{E}_D[L]] = \int \mathbb{E}_D[(y_D(\mathbf{x}) - \mathbb{E}[t/\mathbf{x}])^2]p(\mathbf{x})d\mathbf{x} + \int \text{var}[t/\mathbf{x}]p(\mathbf{x})d\mathbf{x}$$

- Bias-Variance decomposition

$$\mathbb{E}_D[(y_D(\mathbf{x}) - \mathbb{E}[t/\mathbf{x}])^2] = \mathbb{E}_D[(y_D(\mathbf{x}) - \mathbb{E}_D[y_D(\mathbf{x})] + \mathbb{E}_D[y_D(\mathbf{x})] - \mathbb{E}[t/\mathbf{x}])^2]$$

$$= \mathbb{E}_D\Big[\underbrace{\{y_D(\underline{x}) - \mathbb{E}_D[y_D(\underline{x})]\}^2}_{\text{Variance}}\Big] + \mathbb{E}_D\Big[\underbrace{\{\mathbb{E}_D(y_D(\underline{x})) - \mathbb{E}[t/\underline{x}]\}^2}_{\text{Bias}^2}\Big]$$

(Bias)^2 = $\int \{\mathbb{E}_D[y_D(\underline{x})] - \mathbb{E}[t/\underline{x}]\}^2 p(\underline{x})d\underline{x}$

Variance = $\int \mathbb{E}_D\left[\{y_D(\underline{x}) - \mathbb{E}_D[y_D(\underline{x})]\}^2\right] p(\underline{x})d\underline{x}$

Noise = $\int \text{var}[t/\underline{x}]\, p(\underline{x})\, d\underline{x}$

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence & Learning Lab

Example

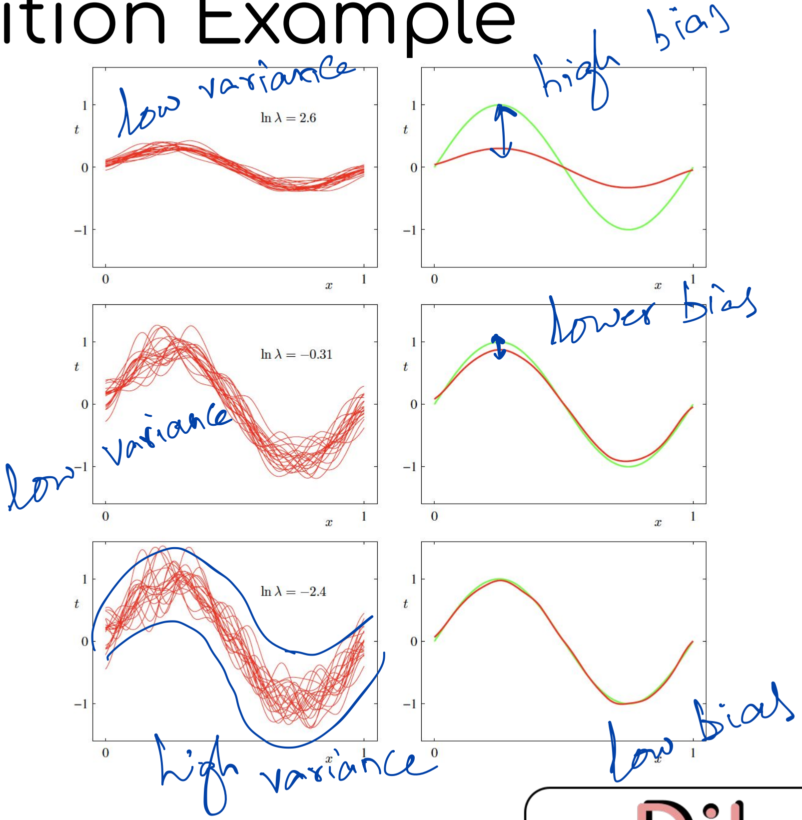
Bias-Variance Decomposition Example

- 100 datasets of size 25
- x ~ U[0, 1]
- t = sin(2πx) + ϵ

$$y^{(l)} = w^{(l)T} \phi(x)$$

l = 1 to 100

$$\mathbb{E}_D[y_D(x)] = \bar{y}(x)$$



Bias-Variance Decomposition Example

(quantifying)

Estimating the bias and variance [since we know the ground truth]

$$(\text{bias})^2 = \int \{\mathbb{E}_D[y_D(x) - \mathbb{E}[t/x]\}^2 p(x)dx \quad = \frac{1}{N}\sum_{i=1}^{N}\left[\bar{y}(x_i) - E(t/x_i)\right]^2$$

$$\frac{1}{L}\sum_{l=1}^{L} y^{(l)}$$

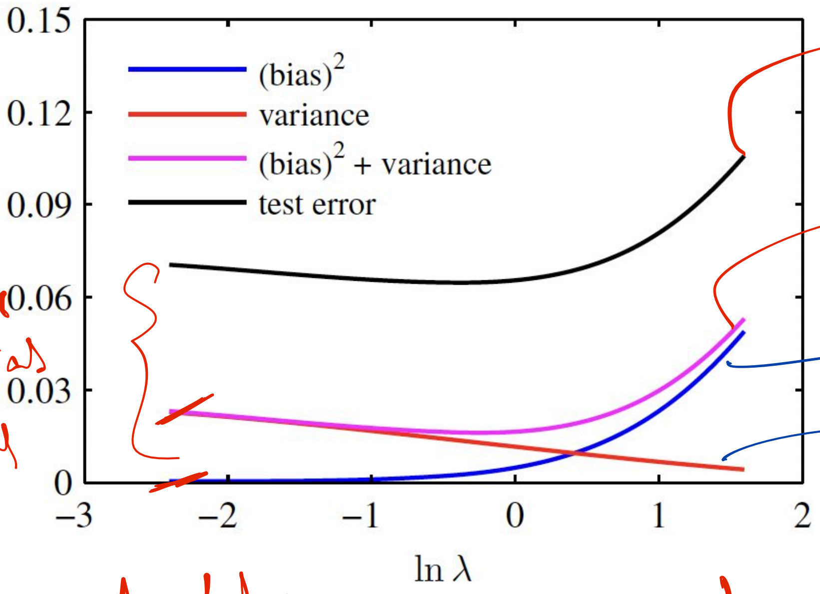
Numerical integration over $\{\underline{x}_1, \underline{x}_2 \ldots \underline{x}_N\}$ monte Carlo approximation

$$\bar{y}(\underline{x}) = \mathbb{E}_D[y_D(\underline{x})] \simeq \frac{1}{L}\sum_{l=1}^{L} y^l(\underline{x})$$

$$\text{variance} = \mathbb{E}_D[\{y_D(x) - \mathbb{E}_D[y_D(x)]\}]^2 p(x)dx$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\left[y^{(l)}(\underline{x}_i) - \bar{y}(\underline{x}_i)\right]^2$$

Bias-Variance Decomposition Example



B+V+N

B+V

Bias

Variance

high var
low bias
overfitting

light → heavy regularization

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence
& Learning Lab

Bias-Variance Decomposition

- In practice - we don't split our dataset to determine the model complexity
    - Large datasets are better
- Bayesian regression!

Rough work

Next
Bayesian Regression