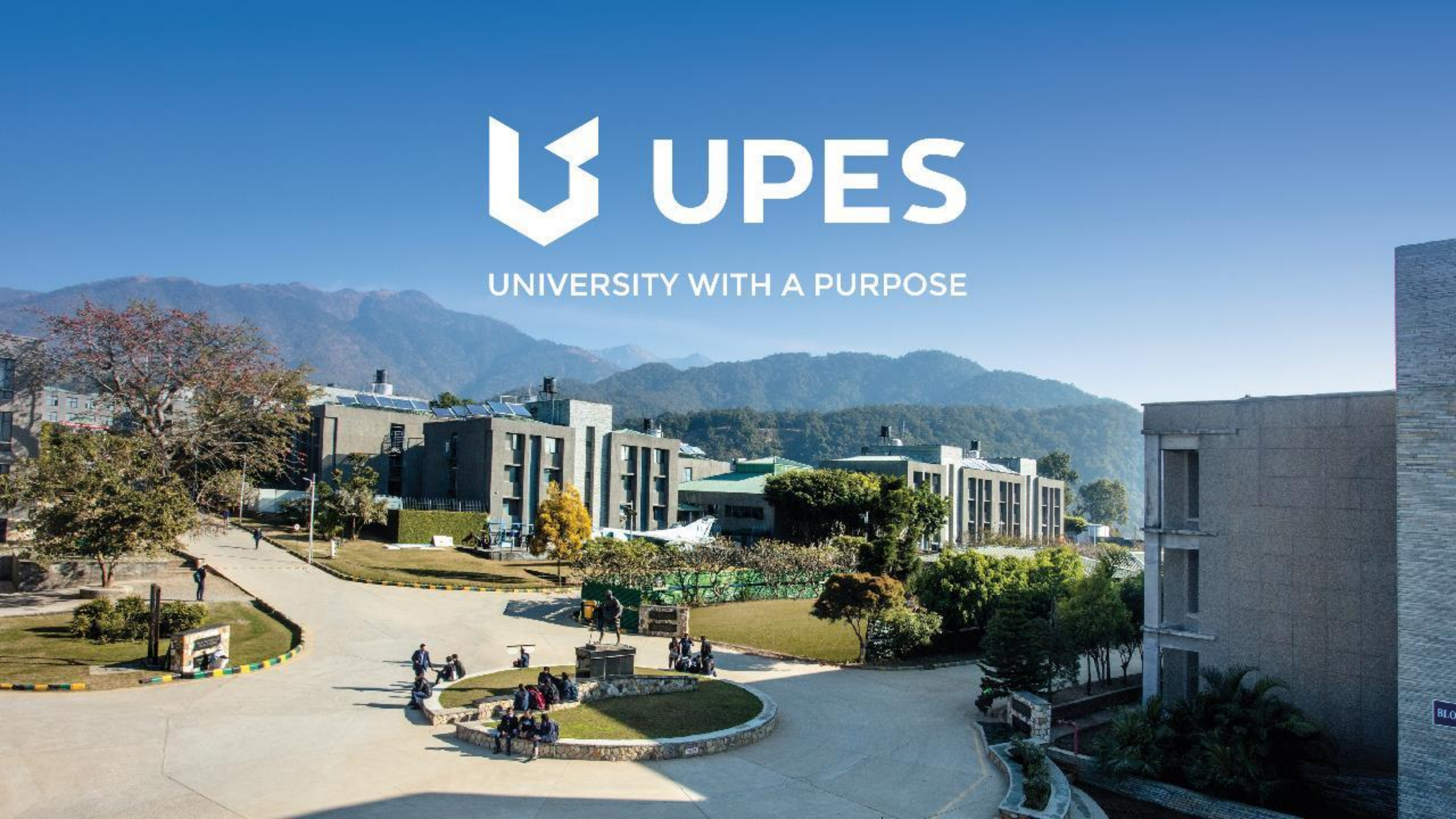
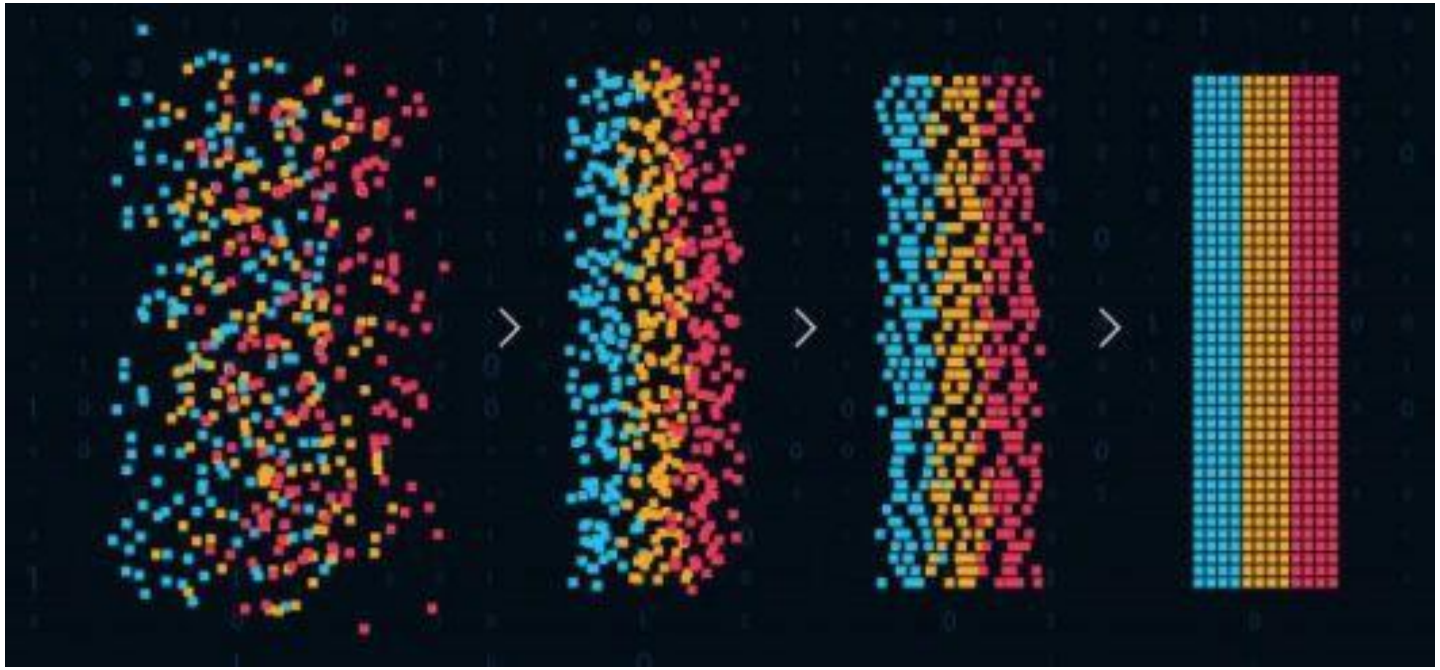

UPES
UNIVERSITY WITH A PURPOSE

Pattern and Anomaly Detection



Source: Edureka

B. Tech., CSE + AI/ML
Dr Gopal Singh Phartiyal
16/09/2021

Bias- Variance

- So far in linear models for regression: We have assumed that the form and number of basis functions are both fixed

Limiting the number of basis functions in order to avoid over-fitting

Vs

limiting the flexibility of the model to capture interesting and important trends in the data

- The introduction of regularization terms can control over-fitting for models with many parameters
- New question: Suitable values of these parameters

Bias- Variance

- Recall the ***expected squared loss***,(not sum of squared error)

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

- where

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x})\,\mathrm{d}t.$$

- The second term of E[L] corresponds to the noise inherent in the random variable t.

- What about the first term?

Bias- Variance

$$\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right]$$
$$= \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}.$$

- Thus we can write

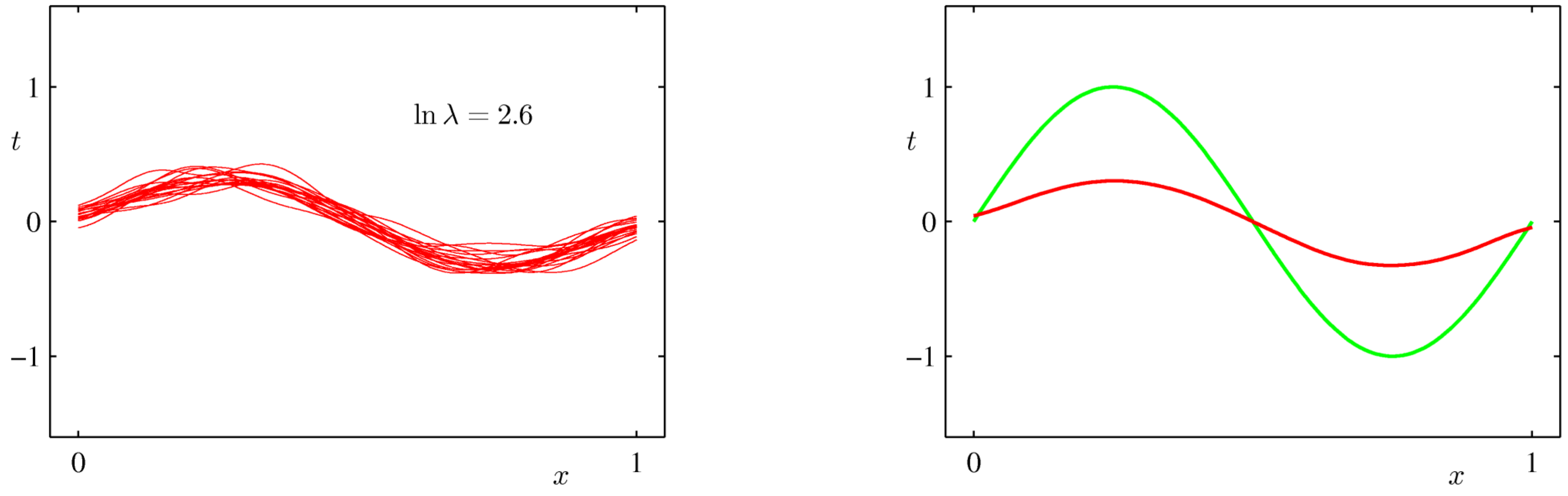
$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

- where

$$\begin{aligned}
(\text{bias})^2 &= \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
\text{variance} &= \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
\text{noise} &= \iint \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t
\end{aligned}$$

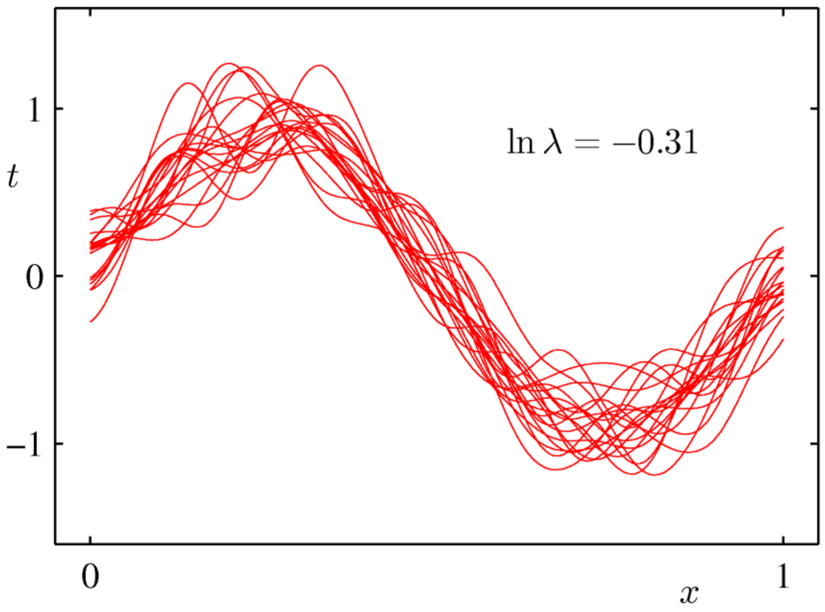
Bias- Variance : Example

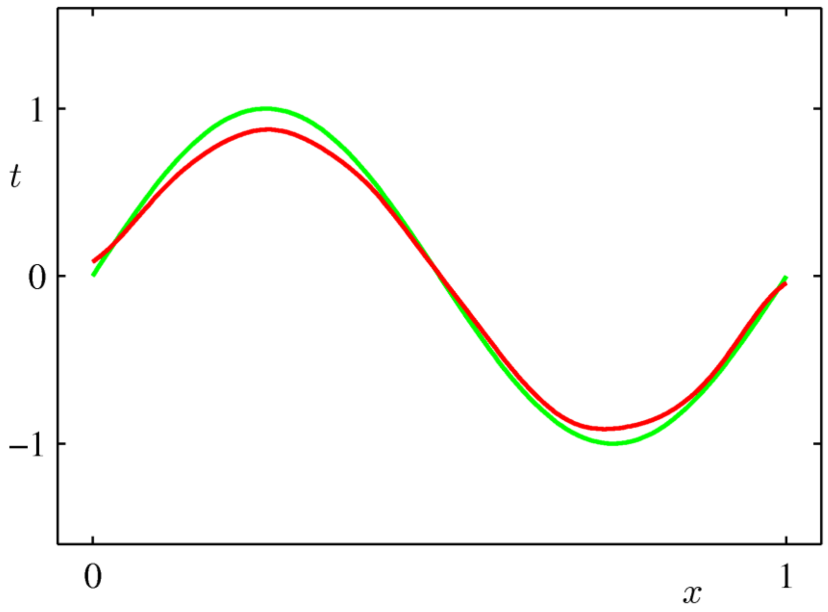
- Example: 25 data sets from the sinusoidal, varying the degree of regularization, λ



Bias- Variance : Example

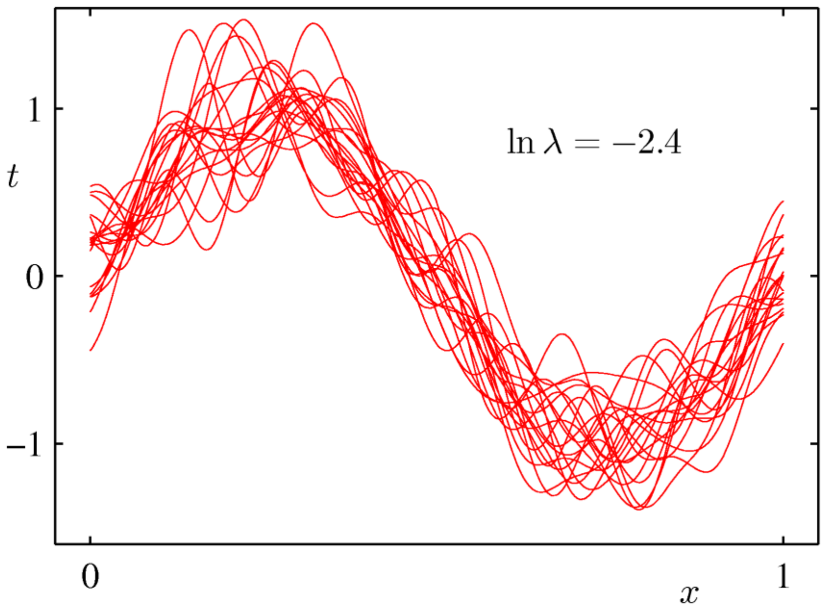
- Example: 25 data sets from the sinusoidal, varying the degree of regularization, λ

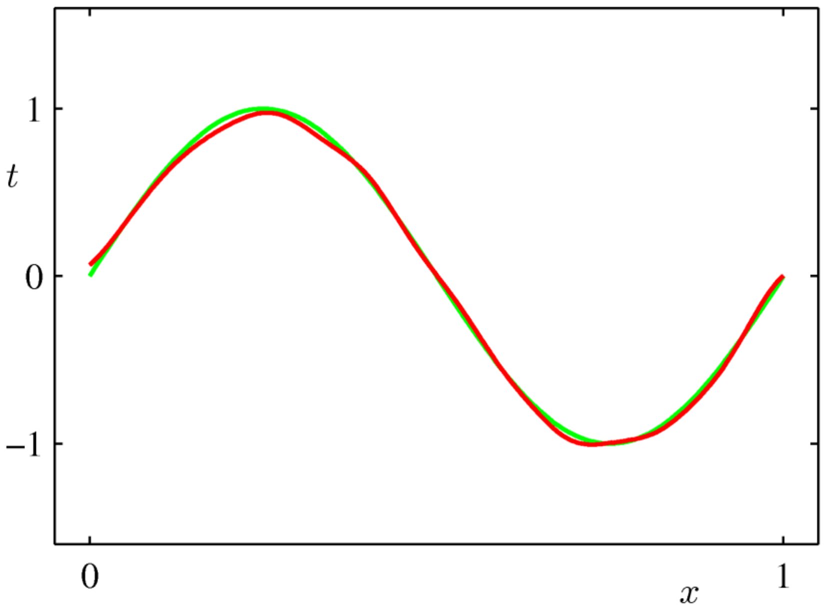




Bias- Variance: Example

- Example: 25 data sets from the sinusoidal, varying the degree of regularization, λ





Bias- Variance: Trade-off

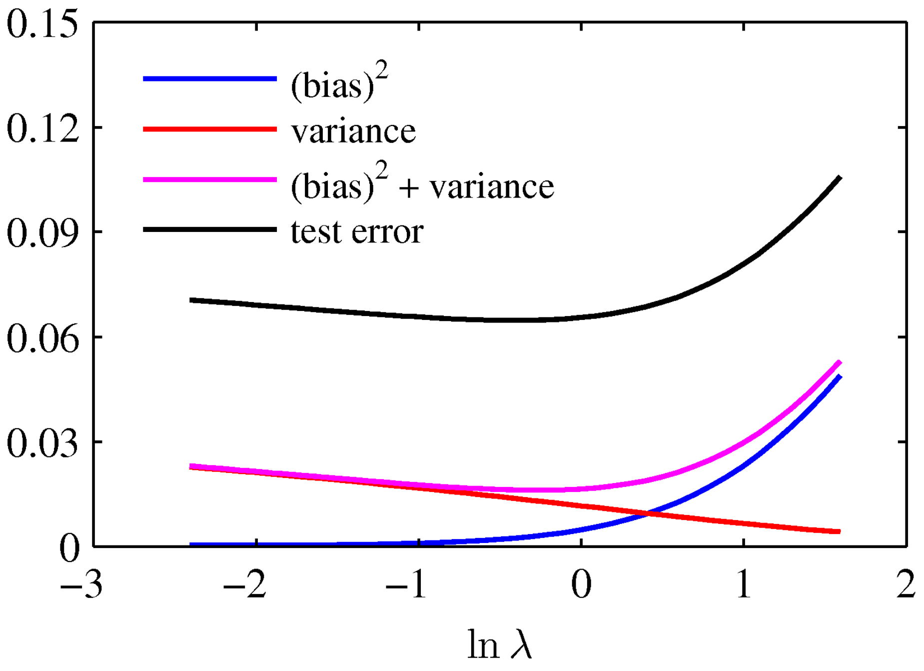
- Example: 25 data sets from the sinusoidal, varying the degree of regularization, λ

•From these plots, we note that an over-regularized model (large λ) will have a high bias, while an under-regularized model (small λ,) will have a high variance.

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

$$(\text{bias})^2 = \frac{1}{N}\sum_{n=1}^{N}\{\overline{y}(x_n) - h(x_n)\}^2$$

$$\text{variance} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\left\{y^{(l)}(x_n) - \overline{y}(x_n)\right\}^2$$



So far in Linear Regression Models

Goal: Find **w?**

- Why linear model?
- Simple linear regression
- Basis functions
- Solving for w using maximum likelihood and least squares
- For multiple output
- Regularize the model (different regularizers)
- Managing over-fitting via the concept of bias-variance decomposition

So which model to choose?

Remember *model selection* and *cross-validation*

Next: Bayesian Linear Regression

- We know - **Posterior = likelihood x prior**
- Likelihood

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}).$$

- Define a conjugate prior over w

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0).$$

•Combining and using results for marginal and conditional Gaussian distributions, gives the posterior *(Refer Chapter-2, Bishop)*

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

- where

$$\begin{aligned}\mathbf{m}_N &= \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}\right) \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}.\end{aligned}$$

Bayesian Linear Regression

- A common choice for the **prior** is (zero-mean isotropic Gaussian)

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

for which

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \mathbf{\Phi}^{\mathrm{T}} \mathbf{t} \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}. \end{aligned}$$

•The w may vary from w_{ML} to w_{prior}

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2} \sum_{n=1}^{N} \{t_n - \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n)\}^2 - \frac{\alpha}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w} + \mathrm{const.}$$

•Maximization of this posterior distribution with respect to w is therefore equivalent to the minimization of the sum-of-squares error

Example

$$y(x, \mathbf{w}) = w_0 + w_1 x.$$

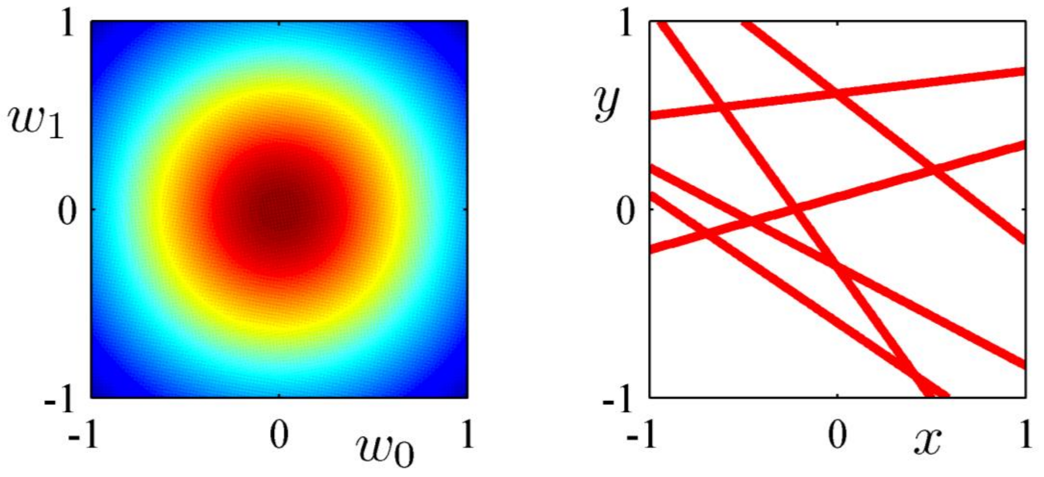
X = U(x| (-1,1), 20 observations

$W_0 = 0.3, \quad w_1 = 0.5$

Noise = Gaussian (sigma = 0.2)

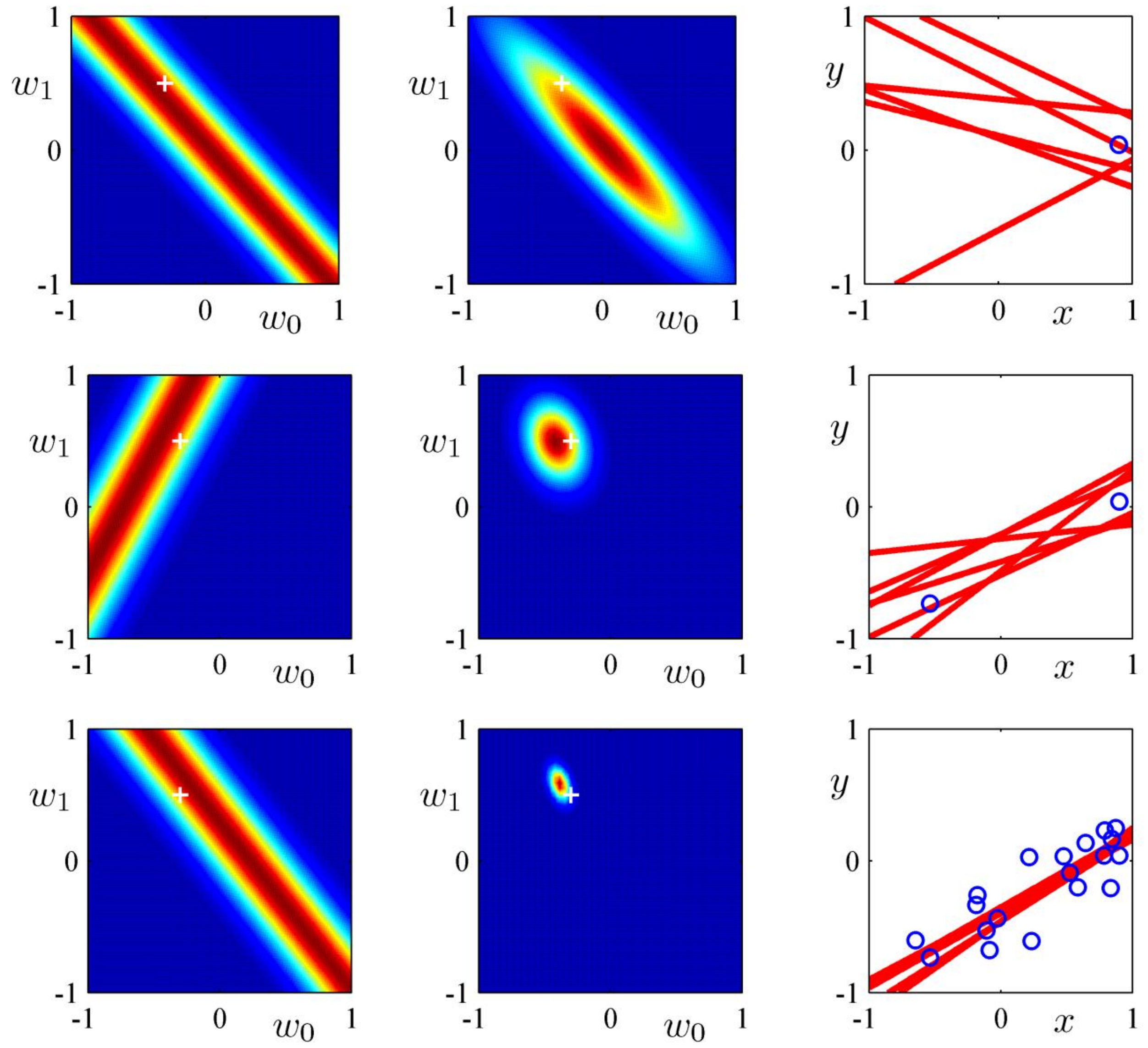
Alpha = 2 for prior

Prior Data Space



0 data points observed

Likelihood Posterior Data Space



Thank You

UPES

UNIVERSITY WITH A PURPOSE