ISTA 421/521 Introduction to Machine Learning

Lecture 9: Maximum Likelihood Uncertainty 2

Clay Morrison

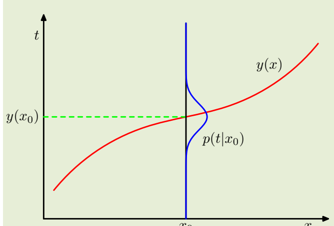
clayton@sista.arizona.edu

Gould-Simpson 819

Phone 621-6609

23 September 2014

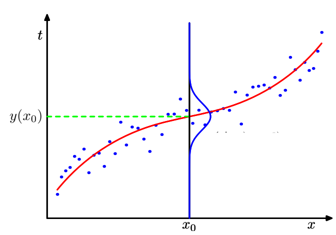
Review



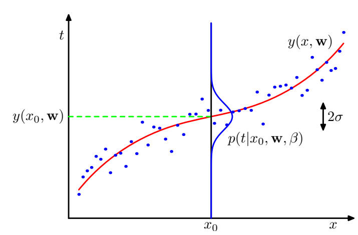
The generating process...

$$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \; ; \; \epsilon \sim \mathcal{N}(0,\sigma^2)$$

$$L = p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n,\mathbf{w},\sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\top\mathbf{x}_n,\sigma^2)$$



... generates data ...



... that we fit a model to

$$p(\hat{\mathbf{t}}|\mathbf{X},\hat{\mathbf{w}},\hat{\sigma^2}) = \prod_{n=1}^{N} p(\hat{t}_n|\mathbf{x}_n,\hat{\mathbf{w}},\hat{\sigma^2}) = \prod_{n=1}^{N} \mathcal{N}(\hat{\mathbf{w}}^\top\mathbf{x}_n,\hat{\sigma^2})$$

prediction

estimated parameters

Maximum Likelihood Estimates of Params

$\hat{\mathbf{w}}$: $\hat{\mathbf{w}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{t}$

$\hat{\sigma}^2$: $\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\hat{\mathbf{w}})$

The MLE is unique

$$\frac{\partial^2 \log L}{\partial\mathbf{w}\partial\mathbf{w}^\top} = -\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}$$

Estimating Uncertainty in Param Estimates via Expected Value

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\ \mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\right\} = \mathbf{w}$$

$$\text{cov}\{\hat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = -\left(\frac{\partial^2 \log L}{\partial\mathbf{w}\partial\mathbf{w}^\top}\right)^{-1}$$

The Fisher Information $\mathcal{I} = \frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X}$

New Predictions:

Variability in Predictions

- We are predicting 2 values:

$$t_{new} \;,\; \sigma^2_{new}$$

$$t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$$ Same solution as minimizing mean squared loss

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{new}\right\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\hat{\mathbf{w}}\right\}^\top \mathbf{x}_{new}\\ &= \mathbf{w}^\top \mathbf{x}_{new}\end{aligned}$$

The ***expected value*** of our prediction is the new data attribute multiplied by the ***true w***

Predicting the Variance of t_{new}

$$\sigma^2_{new} = \mathrm{var}\left\{t_{new}\right\} = \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t^2_{new}\right\} - \left(\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{t_{new}\right\}\right)^2$$

Substitute $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$

$$\begin{aligned}\mathrm{var}\{t_{\mathrm{new}}\} &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{(\widehat{\mathbf{w}}^\mathsf{T}\mathbf{x}_{\mathrm{new}})^2\right\} - (\mathbf{w}^\mathsf{T}\mathbf{x}_{\mathrm{new}})^2\\ &= \mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{x}_{\mathrm{new}}^\mathsf{T}\widehat{\mathbf{w}}\widehat{\mathbf{w}}^\mathsf{T}\mathbf{x}_{\mathrm{new}}\right\} - \mathbf{x}_{\mathrm{new}}^\mathsf{T}\mathbf{w}\mathbf{w}^\mathsf{T}\mathbf{x}_{\mathrm{new}}.\end{aligned}$$

Substitute $\widehat{\mathbf{w}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$

$$\mathrm{var}\{t_{\mathrm{new}}\} = \mathbf{x}_{\mathrm{new}}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\mathsf{T}\right\}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}} - \mathbf{x}_{\mathrm{new}}^\mathsf{T}\mathbf{w}\mathbf{w}^\mathsf{T}\mathbf{x}_{\mathrm{new}}$$

On slide 22 of lec 8, in the derivation of the covariance of $\hat{\boldsymbol{w}}$, we identified

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}\mathbf{t}^\top\right\} = \mathbf{X}\mathbf{w}\mathbf{w}^\top\mathbf{X}^\top - \sigma^2\mathbf{I}$$

$$\begin{aligned}\mathrm{var}\{t_{\mathrm{new}}\} &= \mathbf{x}_{\mathrm{new}}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}(\sigma^2\mathbf{I} + \mathbf{X}\mathbf{w}\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T})\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}} - \mathbf{x}_{\mathrm{new}}^\mathsf{T}\mathbf{w}\mathbf{w}^\mathsf{T}\mathbf{x}_{\mathrm{new}}\\ &= \sigma^2\mathbf{x}_{\mathrm{new}}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}} + \mathbf{x}_{\mathrm{new}}^\mathsf{T}\mathbf{w}\mathbf{w}^\mathsf{T}\mathbf{x}_{\mathrm{new}} - \mathbf{x}_{\mathrm{new}}^\mathsf{T}\mathbf{w}\mathbf{w}^\mathsf{T}\mathbf{x}_{\mathrm{new}}\\ &= \sigma^2\mathbf{x}_{\mathrm{new}}^\mathsf{T}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}}.\end{aligned}$$

Recall: $\mathrm{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = -\left(\frac{\partial^2 \log L}{\partial\mathbf{w}\partial\mathbf{w}^\mathsf{T}}\right)^{-1}$

So, could be written

$$\sigma^2_{\mathrm{new}} = \mathbf{x}_{\mathrm{new}}^\mathsf{T}\mathrm{cov}\{\widehat{\mathbf{w}}\}\mathbf{x}_{\mathrm{new}}$$

Prediction Summary

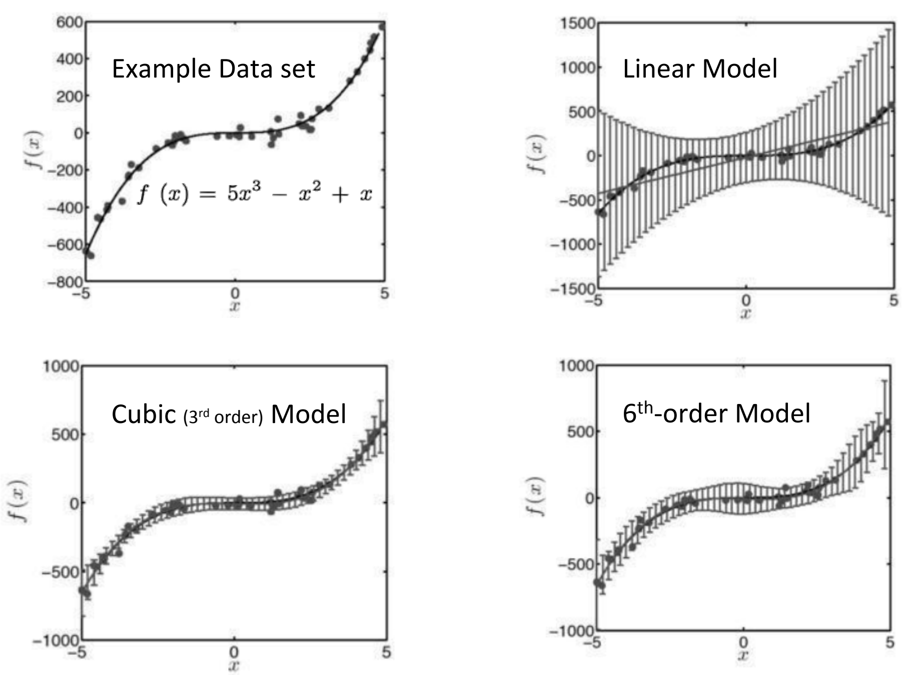
$$t_{\text{new}} = \mathbf{x}_{\text{new}}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{t} = \mathbf{x}_{\text{new}}^{\mathsf{T}}\widehat{\mathbf{w}}$$

$$\sigma_{\text{new}}^2 = \sigma^2 \mathbf{x}_{\text{new}}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{x}_{\text{new}}$$

We estimate this from the data: $\hat{\sigma}^2$

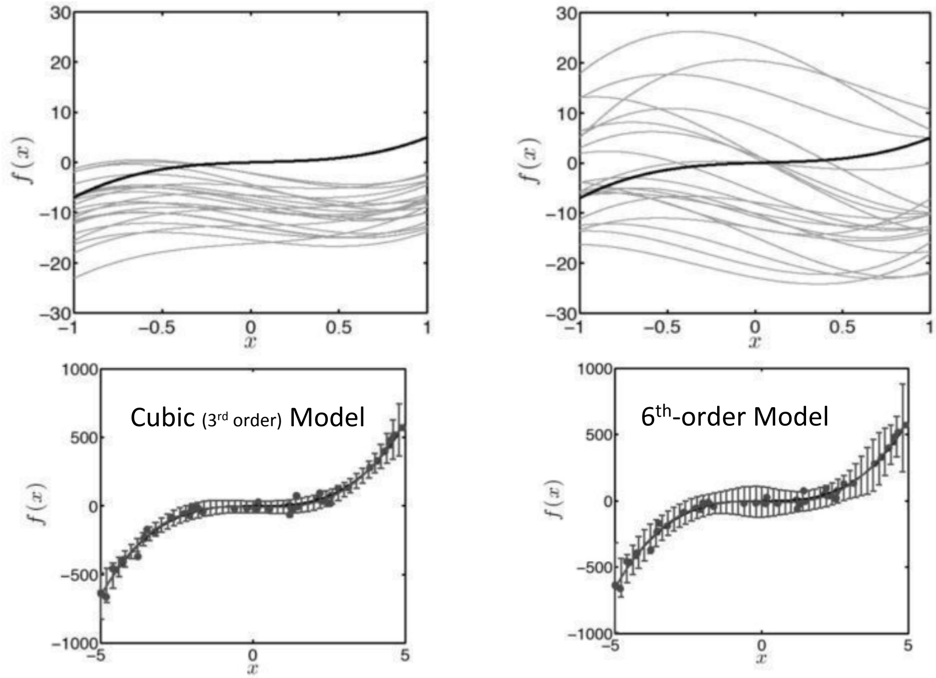
$t_{\text{new}} \pm \sigma_{\text{new}}^2$

Plotting Predictive Error Bars

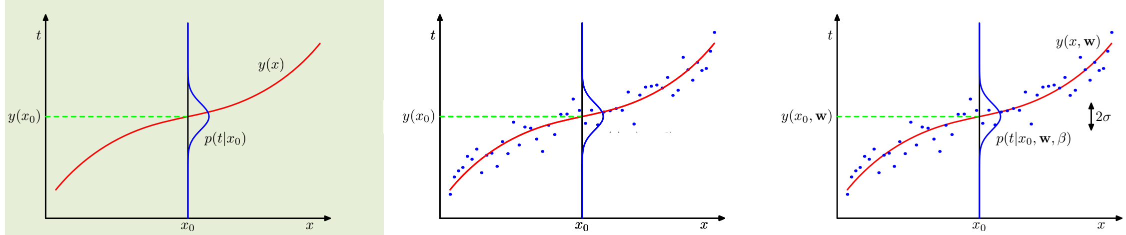


$t_{\mathrm{new}} \pm \sigma^2_{\mathrm{new}}$

Plotting Predictive Error Bars



Review



The generating process… ... generates data ... ... that we fit a model to

$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \;\; ; \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\top \mathbf{x}_n, \sigma^2)$$

$$p(\hat{\mathbf{t}}|\mathbf{X}, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} p(\hat{t}_n|\mathbf{x}_n, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} \mathcal{N}(\hat{\mathbf{w}}^\top \mathbf{x}_n, \hat{\sigma^2})$$

prediction

estimated parameters

Maximum Likelihood Estimates of Params

$\hat{\mathbf{w}}$: $\widehat{\mathbf{w}} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{t}$

$\hat{\sigma}^2$: $\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\mathsf{T}\mathbf{t} - \mathbf{t}^\mathsf{T}\mathbf{X}\widehat{\mathbf{w}})$

The MLE is unique

$\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} = -\frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}$

Estimating Uncertainty in Param Estimates via Expected Value

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\ \mathbf{w}, \sigma^2)} \left\{\widehat{\mathbf{w}}\right\} = \mathbf{w}$

The Fisher Information $\mathcal{I} = \frac{1}{\sigma^2}\mathbf{X}^\mathsf{T}\mathbf{X}$

$\mathrm{cov}\{\widehat{\mathbf{w}}\} = \sigma^2(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = -\left(\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}}\right)^{-1}$

New Predictions: $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$

$\sigma^2_{\mathrm{new}} = \sigma^2 \mathbf{x}^\mathsf{T}_{\mathrm{new}}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{x}_{\mathrm{new}}$

$\sigma^2_{\mathrm{new}} = \mathbf{x}^\mathsf{T}_{\mathrm{new}}\mathrm{cov}\{\widehat{\mathbf{w}}\}\mathbf{x}_{\mathrm{new}}$

Quantifying the Uncertainty in our Estimate of $\hat{\sigma}^2$

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\widehat{\mathbf{w}})$$

$$\begin{aligned}
\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} &= \frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\widehat{\mathbf{w}}\right\} \\
&= \frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}\right\} \\
&= \frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}^\top\mathbf{t}\right\} - \frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}\right\}
\end{aligned}$$

We have seen the expectation of $\mathbf{t}\mathbf{t}^\top$ … but not $\mathbf{t}^\top\mathbf{t}$

$$\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\mathbf{E}_{p(\mathbf{t})}\left\{\mathbf{t}^\top\mathbf{A}\mathbf{t}\right\} = \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top\mathbf{A}\boldsymbol{\mu}$$

Quantifying the Uncertainty in our Estimate of $\hat{\sigma}^2$

- We will need the matrix ***trace***

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1D} \\ A_{21} & A_{22} & \cdots & A_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ A_{D1} & A_{D2} & \cdots & A_{DD} \end{bmatrix}$$

$$\mathrm{Tr}(\mathbf{A}) = \sum_{d=1}^{D} A_{dd}.$$

Other Properties/Identities

$$\mathrm{Tr}(\mathbf{I}_D) = \sum_{d=1}^{D} 1 = D$$

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$$

$$\mathrm{Tr}(a) = a$$

$$\mathrm{Tr}(\mathbf{w}^\top\mathbf{w}) = \mathbf{w}^\top\mathbf{w}$$

- The trace of a matrix is the sum of the (complex) eigenvalues, and is invariant with respect to change of basis.
- Geometrically: The trace can be interpreted as the infinitesimal change in volume (as the derivative of the determinant)

Quantifying the Uncertainty in our Estimate of $\hat{\sigma}^2$

$\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$\mathbf{E}_{p(\mathbf{t})}\left\{\mathbf{t}^\top \mathbf{A} \mathbf{t}\right\} = \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu}$

$\mathrm{Tr}(\mathbf{A}) = \sum_{d=1}^{D} A_{dd}.$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}^\top\mathbf{t}\right\} - \frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}\right\}$$

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} &= \frac{1}{N}\left(\mathrm{Tr}(\sigma^2\mathbf{I_N}) + \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}\right) \\ &- \frac{1}{N}\left(\mathrm{Tr}(\sigma^2\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) + \mathbf{w}^\top\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{w}\right)\end{aligned}$$

$\mathrm{Tr}(\sigma^2\mathbf{A}) = \sigma^2\mathrm{Tr}(\mathbf{A})$ and $\mathrm{Tr}(\mathbf{I_N}) = N$

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} &= \sigma^2 + \frac{1}{N}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} - \frac{\sigma^2}{N}\mathrm{Tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) - \frac{1}{N}\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w} \\ &= \sigma^2 - \frac{\sigma^2}{N}\mathrm{Tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top) \\ &= \sigma^2\left(1 - \frac{1}{N}\mathrm{Tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\right).\end{aligned}$$

$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$

$$\begin{aligned}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} &= \sigma^2\left(1 - \frac{1}{N}\mathrm{Tr}((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X})\right) \\ &= \sigma^2\left(1 - \frac{1}{N}\mathrm{Tr}(\mathbf{I}_D)\right) \\ &= \sigma^2\left(1 - \frac{D}{N}\right),\end{aligned}$$

D is the number of attributes (the number of columns in $\mathbf{X}$)

Quantifying the Uncertainty in our Estimate of $\hat{\sigma}^2$

$\mathbf{t} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$\mathbf{E}_{p(\mathbf{t})}\left\{\mathbf{t}^\top \mathbf{A} \mathbf{t}\right\} = \mathrm{Tr}(\mathbf{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu}$

$\mathrm{Tr}(\mathbf{A}) = \sum_{d=1}^{D} A_{dd}.$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}^\top\mathbf{t}\right\} - \frac{1}{N}\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\mathbf{t}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}\right\}$$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right)$$

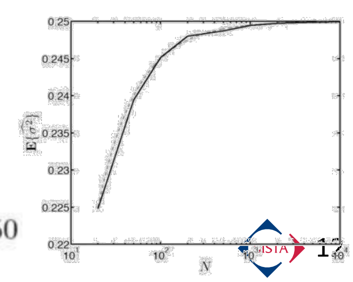
When D < N (that is, the number of attributes we measure for each data point is *smaller* than the number of data points), then our estimates of the variance will, on average, be lower than the true variance.

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} < \sigma^2$$

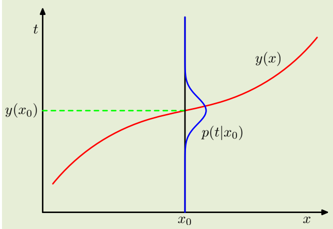
Unlike the estimate for $\hat{\mathbf{w}}$, the MLE for $\hat{\sigma}^2$ is ***biased***.

$D = 2$ and $N = 20$

$$\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\} = \sigma^2\left(1 - \frac{D}{N}\right) = 0.25\left(1 - \frac{2}{20}\right) = 0.2250$$



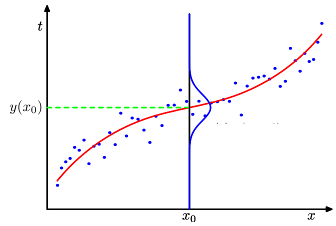
Review



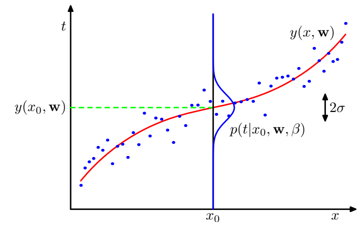
The generating process...

$t_n = \mathbf{w}^\top \mathbf{x}_n + \epsilon_n \;\; ; \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$

$$L = p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{w}^\top \mathbf{x}_n, \sigma^2)$$



... generates data ...



... that we fit a model to

$$p(\hat{\mathbf{t}}|\mathbf{X}, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} p(\hat{t}_n|\mathbf{x}_n, \hat{\mathbf{w}}, \hat{\sigma^2}) = \prod_{n=1}^{N} \mathcal{N}(\hat{\mathbf{w}}^\top \mathbf{x}_n, \hat{\sigma^2})$$

prediction

estimated parameters

Maximum Likelihood Estimates of Params

$\hat{\mathbf{w}}$: $\widehat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{t}$

$\hat{\sigma}^2$: $\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t}^\top \mathbf{t} - \mathbf{t}^\top \mathbf{X} \widehat{\mathbf{w}})$

The MLE is unique

$\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\top} = -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X}$

Estimating Uncertainty in Param Estimates via Expected Value

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)} \{\widehat{\mathbf{w}}\} = \mathbf{w}$

$\mathbf{E}_{p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)} \{\widehat{\sigma^2}\} = \sigma^2 \left(1 - \frac{D}{N}\right)$

The Fisher Information $\mathcal{I} = \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{X}$

$\text{cov}\{\widehat{\mathbf{w}}\} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} = -\left(\frac{\partial^2 \log L}{\partial \mathbf{w} \partial \mathbf{w}^\top}\right)^{-1}$

New Predictions: $t_{new} = \hat{\mathbf{w}}^\top \mathbf{x}_{new}$

$\sigma^2_{\text{new}} = \sigma^2 \mathbf{x}_{\text{new}}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_{\text{new}}$

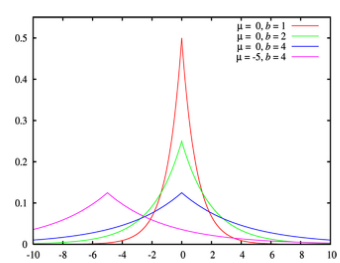
$\sigma^2_{\text{new}} = \mathbf{x}_{\text{new}}^\top \text{cov}\{\widehat{\mathbf{w}}\} \mathbf{x}_{\text{new}}$

MLE Beyond the Gaussian

- Other useful probability distributions have analytic MLE solutions.

- The MLE of **Laplace** i.i.d. variables.

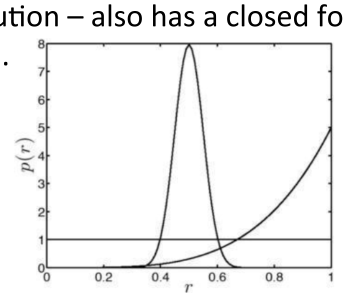
$$f(x|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right) = \frac{1}{2b} \begin{cases} \exp\left(-\frac{\mu - x}{b}\right) & \text{if } x < \mu \\ \exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu \end{cases}$$



- μ = "location" parameter, b = "diversity"

- Another univariate case is the **beta** distribution – also has a closed form, with lots of flexibility in the shape/location.

$$p(r) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha - 1} (1 - r)^{\beta - 1}$$

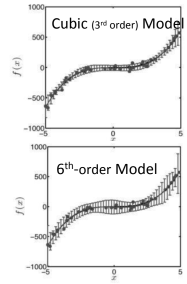


MLE and Model Selection

$$\log L = -\frac{1}{N}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\top \mathbf{x}_n)^2$$

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n-1}^{N}(t_n - \mathbf{x}^\top\hat{\mathbf{w}})^2 = \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\hat{\mathbf{w}})$$

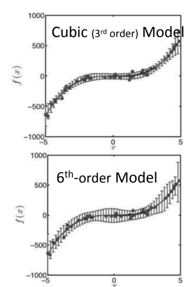


MLE Prefers Complex Models

$$\log L = -\frac{1}{N}\log 2\pi - N\log\sigma - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \mathbf{w}^\top \mathbf{x}_n)^2$$

$$\hat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{t}$$

$$\widehat{\sigma^2} = \frac{1}{N}\sum_{n-1}^{N}(t_n - \mathbf{x}^\top\hat{\mathbf{w}})^2 = \frac{1}{N}(\mathbf{t}^\top\mathbf{t} - \mathbf{t}^\top\mathbf{X}\hat{\mathbf{w}})$$



Plug in $\widehat{\sigma^2}$ to the log likelihood:

$$\begin{aligned}\log L &= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\widehat{\sigma^2} - \frac{1}{2\widehat{\sigma^2}}N\widehat{\sigma^2} \\ &= -\frac{N}{2}(1 - \log 2\pi) - \frac{N}{2}\log\widehat{\sigma^2}\end{aligned}$$

Making $\hat{\sigma}^2$ smaller makes log L larger.

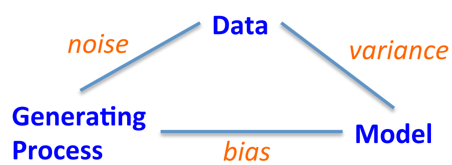
Bad news: Increasing the model complexity will *decrease* the variance!

Bottom line: Unfortunately, we can't use MLE to do model selection
But, with a particular model, MLE will choose the parameters that make the data have the highest overall likelihood under the model.

Bias-Variance Tradeoff

- **Bias**: the systematic mismatch between our model and the process that generated the data.
  - Too simple a model == too high a bias (underfitting)
  - Too complex a model (too many degrees of freedom) == too low a bias (overfitting)
- **Variance**: Squared error between model and data

- Imagine we had the *true* distribution that generated the data; we could compute and compare the expected value of the squared error between estimated parameter values and the true values.
- We would like this value to be as small as possible.



The Bias-Variance Decomposition (1)

$$\mathbf{E}_{p(x)}\left\{f(x)\right\} = \int f(x)p(x)\,dx$$

Sorry to switch Expectation notation on you…
← to the left is from our class book
The notation below and the next couple of slides comes from Bishop (2007)

- Recall the *expected squared loss*,

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

- 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?

The Bias-Variance Decomposition (2)

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

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

- Suppose we were given multiple data sets, each of size *N*. Any particular data set, *D*, will give a particular function *y*(*x*;*D*). We then have

$$\begin{aligned}
&\{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2 \\
&\quad = \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\
&\quad = \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\
&\qquad + 2\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}.
\end{aligned}$$

The Bias-Variance Decomposition (3)

- Taking the expectation over *D* yields

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

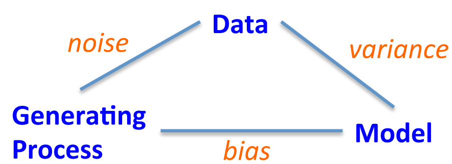
The Bias-Variance Decomposition (4)

- Thus we can write

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

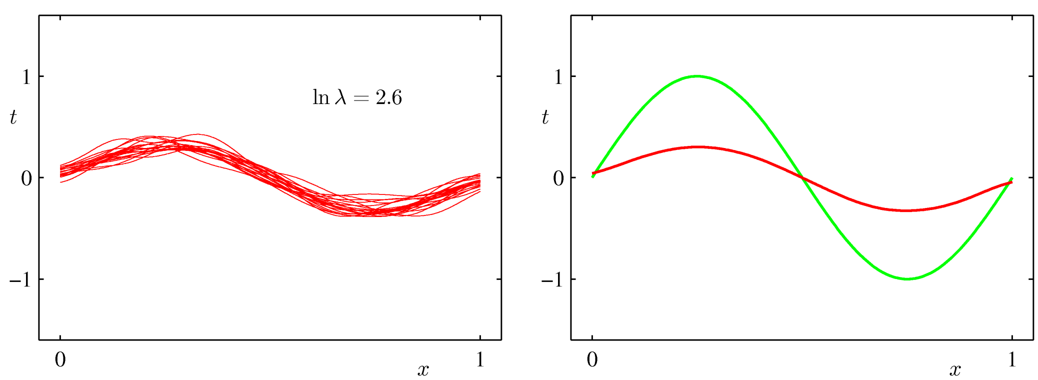
- where

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



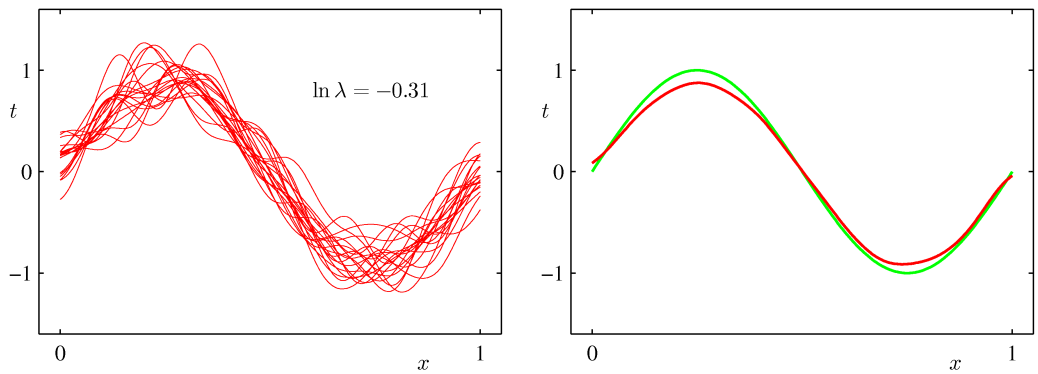
The Bias-Variance Decomposition (5)

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



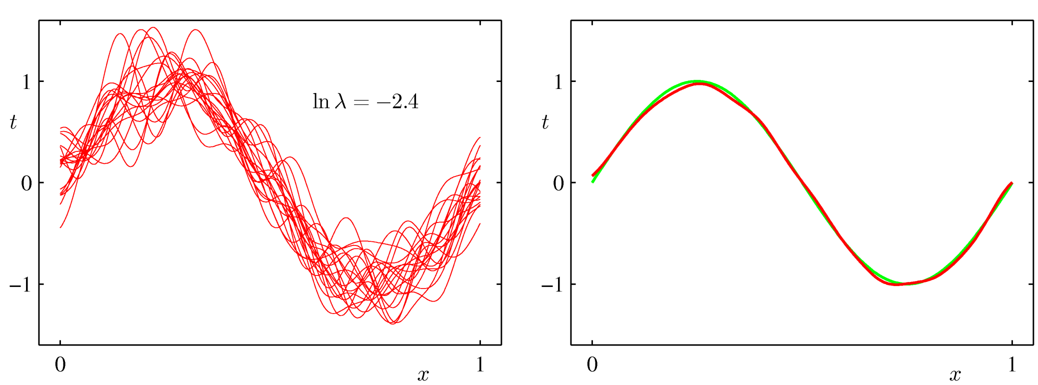
The Bias-Variance Decomposition (6)

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



The Bias-Variance Decomposition (7)

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



The Bias-Variance Trade-off

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.