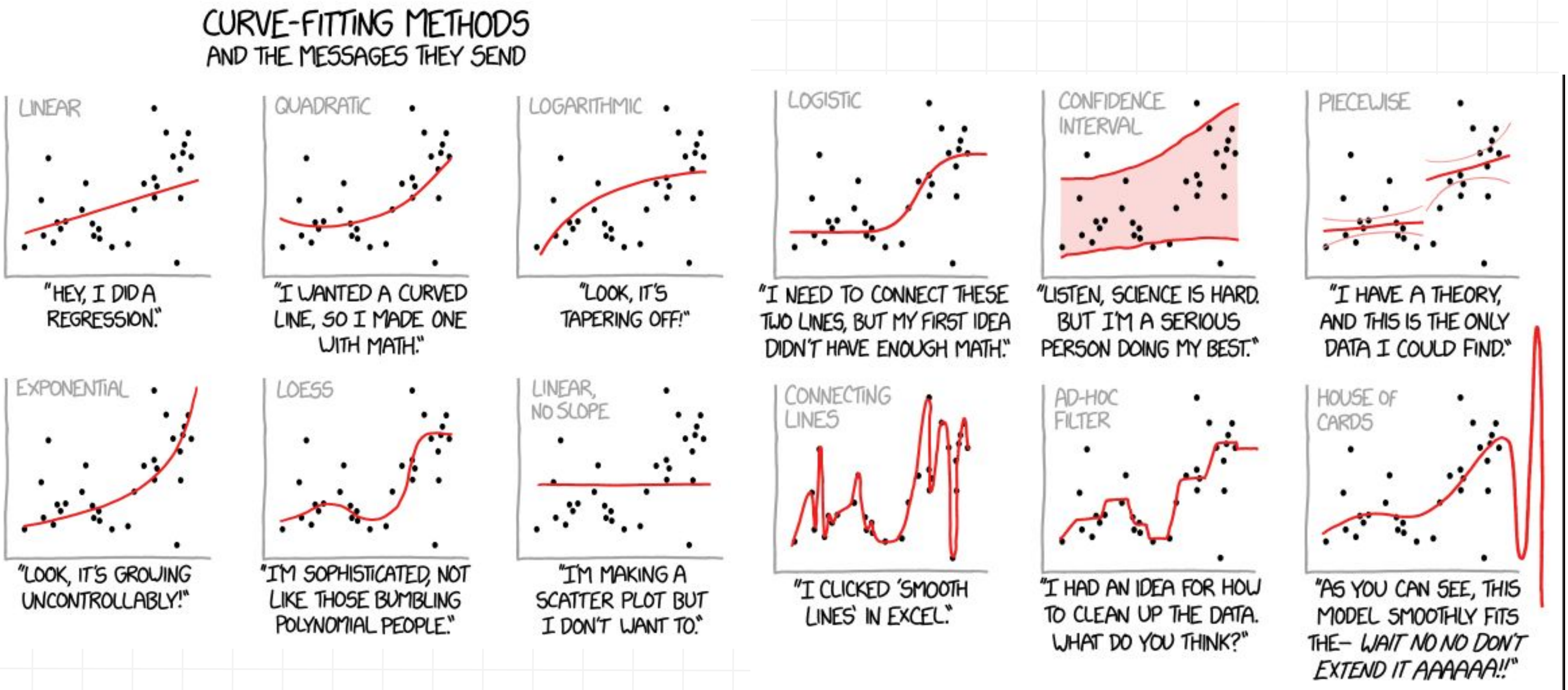
https://xkcd.com/2048/



Linear Regression (Linear models for regression)

Why linear models?

Input data, features, basis functions

Maximum likelihood and least squares

Geometric intuition

Regularised least squares

Multiple outputs

Bias-variance decomposition

The relation between MLE and least squares, Lagrange multipliers, multiple ways of looking at linear models.

Why linear models? - it saves lives

The five criteria of the Apgar score:[3]

	Score of 0	Score of 1	Score of 2	Component of backronym
Skin color	blue or pale all over	blue at extremities, body pink (acrocyanosis)	no cyanosis body and extremities pink	**A**ppearance
Pulse rate	absent	< 100 beats per minute	≥ 100 beats per minute	**P**ulse
Reflex irritability grimace	no response to stimulation	grimace on suction or aggressive stimulation	cry on stimulation	**G**rimace
Muscle Tone	none	some flexion	flexed arms and legs that resist extension	**A**ctivity
Respiratory effort	absent	weak, irregular, gasping	strong, robust cry	**R**espiration

Virginia Apgar, creator of the Apgar score

Purpose method to summarize newborn's health

What do the Apgar scores mean?

A score of 7 or more is normal. A score of 6 or less at 1 minute and a score of 7 or more at 5 minutes is also normal. However, a score below 7 in the second test at 5 minutes is considered low.

If your baby's score was low in the first Apgar test and hasn't improved in the second test at 5 minutes, or there are other concerns, the doctors and nurses will closely monitor your baby and continue any necessary medical care.

Image and table from wikipedia
https://www.pregnancybirthbaby.org.au/apgar-score

Everything should be made as simple as possible, but not simpler. -- Albert Einstein

Occam's razor, also spelled **Ockham's razor**, also called **law of economy** or **law of parsimony**, principle stated by the Scholastic philosopher William of Ockham (1285–1347/49) that *pluralitas non est ponenda sine necessitate*, "plurality should not be posited without necessity." The principle gives precedence to simplicity: of two competing theories, the simpler explanation of an entity is to be preferred. The principle is also expressed as "Entities are not to be multiplied beyond necessity."

https://www.britannica.com/topic/Occams-razor

SNAP-II and SNAPPE II (Score for Neonatal Acute Physiology and SNAP Perinatal Extension)

Variables	Values	Points
Mean Blood Pressure		0
Lowest temperature		0
$P0_2$ (mmHg) / FIO_2 (%)		0
Lowest serum pH		0
Multiple seizures		0
Urine output (mL/kg.h)		0
	SNAP II : 0	
Apgar score		
Birth weight		
Small for gestational age (help)		

SNAPPE II : 0 In- hospital mortality : see below Data are collected within t NICU

Clear

Ref: D K. Richardson et al . SNAP-II and SNAPPE-II: Simplified newborn illness severity and morta 2001; 138: 92-100

The Lancet, 2003

Research Letters

CRIB II: an update of the clinical risk index for babies score

Dr Gareth Parry PhD [a], Janet Tucker PhD [b], William Tarnow-Mordi MRCP [c], for the UK Neonatal Staffing Study Collaborative Group [*]

The clinical risk index for babies (CRIB) score is a risk-adjustment instrument used worldwide in neonatal intensive care.[1] It was developed with data relating to babies born at less than 31 weeks' gestation, or 1500 g birthweight or lower, admitted for neonatal intensive care between 1988 and 1990. The appropriateness of CRIB with contemporary data has been questioned, since the score may now be poorly calibrated with mortality after neonatal intensive care, in which case it might be no better in prediction of mortality than birthweight or gestation alone. Furthermore, CRIB includes, as one of two measures of severity of illness, fraction of inspired oxygen (FiO_2), which is not a true physiological measure because it is determined by the care team. CRIB also includes data up to 12 h after admission, thus potentially introducing early treatment bias.

https://sfar.org/scores2/snap22.php
https://www.sciencedirect.com/science/article/abs/pii/S0140673603133971

Linear models and likelihood

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

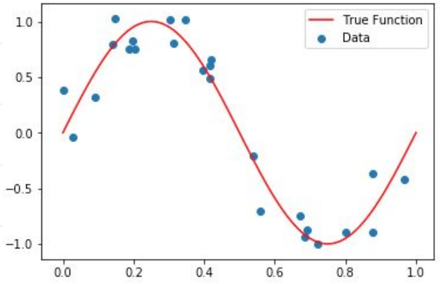
$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

- Linear combination of fixed nonlinear basis functions
- parameter $\mathbf{w} = (w_0, \ldots, w_{M-1})^T$
- basis functions $\boldsymbol{\phi}(\mathbf{x}) = (\phi_0(\mathbf{x}), \ldots, \phi_{M-1}(\mathbf{x}))^T$
- convention $\phi_0(\mathbf{x}) = 1$
- w_0 is the bias parameter

Why linear regression?

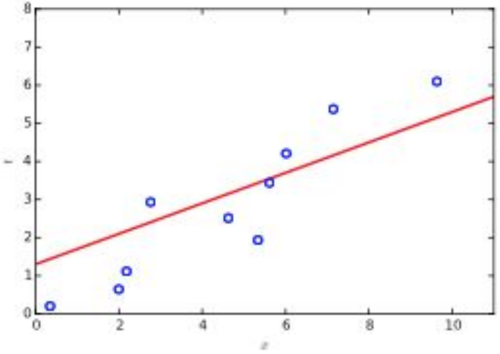
- Analytic solution when minimising sum of squared errors
- Well understood statistical behaviour
- Efficient algorithms exist for convex losses and regularizers

But what if ?



Conventions in [Bishop 2006]

$\mathbf{x}$



$\phi_j(\mathbf{x})$

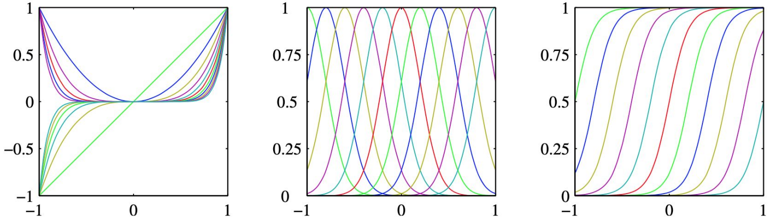
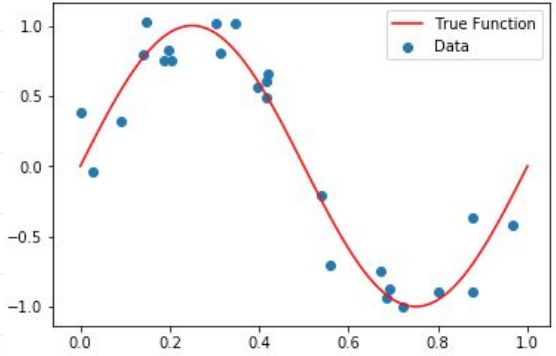


Figure 3.1 Examples of basis functions, showing polynomials on the left, Gaussians of the form (3.4) in the centre, and sigmoidal of the form (3.5) on the right.

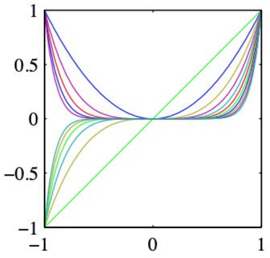


Polynomial Basis Functions

- Scalar input variable x

$$\phi_j(x) = x^j$$

- Limitation : Polynomials are global functions of the input variable x so the learned function will extrapolate poorly

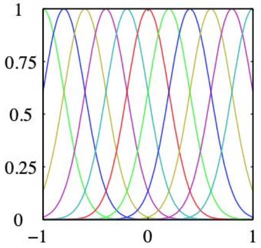


'Gaussian' Basis Functions

- Scalar input variable x

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$

- Not a probability distribution.
- No normalisation required, taken care of by the model parameters w.
- Well behaved away from the data (though pulled to zero)



Sigmoidal Basis Functions

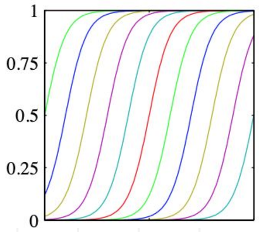
- Scalar input variable x

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$$

where $\sigma(a)$ is the logistic sigmoid function defined by

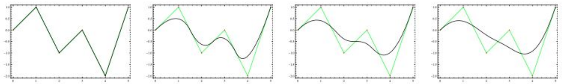
$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

- $\sigma(a)$ is related to the hyperbolic tangent $\tanh(a)$ by $\tanh(a) = 2\sigma(a) - 1$.



Other Basis Functions

- Fourier Basis : each basis function represents a specific frequency and has infinite spatial extent.
- Wavelets : localised in both space and frequency (also mutually orthogonal to simplify appliciation).
- Splines (piecewise polynomials restricted to regions of the input space; additional constraints where pieces meet, e.g. smoothness constraints $\rightarrow$ conditions on the derivatives).



Linear Splines | Quadratic Splines | Cubic Splines | Quartic Splines

Approximate the points $\{(0,0), (1,1), (2,-1), (3,0), (4,-2), (5,1)\}$ by different splines.

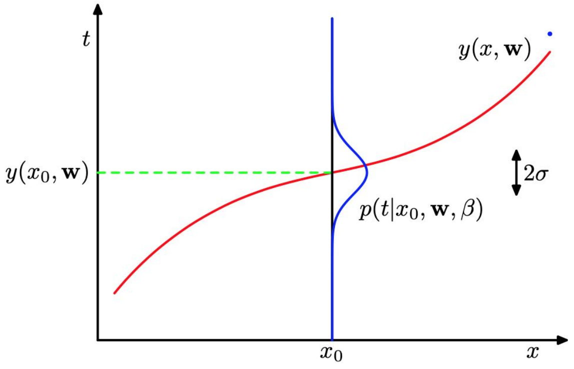
Linear models and likelihood

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \qquad (3.7)$$

Fig 1.16



$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right) \qquad (1.60)$$ also (3.8)

Computing the likelihood

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right) \tag{1.60}$$

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

$$\begin{aligned}
\ln p(\mathbf{t} \,|\, \mathbf{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n \,|\, \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \\
&= \sum_{n=1}^{N} \ln \left(\sqrt{\frac{\beta}{2\pi}} \exp\left\{ -\frac{\beta}{2} (t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n))^2 \right\} \right)
\end{aligned}$$

= ... (see next page)

Computing the likelihood

$$\begin{aligned}\ln p(\mathbf{t}\,|\,\mathbf{w},\beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n\,|\,\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\ &= \sum_{n=1}^{N} \ln\left(\sqrt{\frac{\beta}{2\pi}}\exp\left\{-\frac{\beta}{2}(t_n-\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n))^2\right\}\right) \\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})\end{aligned} \tag{3.11}$$

the sum-of-squares error function is defined by

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2. \tag{3.12}$$

Claim: $\arg\max_{\mathbf{w}} \ln p(\mathbf{t}\,|\,\mathbf{w},\beta) \rightarrow \arg\min_{\mathbf{w}} E_D(\mathbf{w})$

… first rewrite the error function

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2. \qquad (3.12)$$

$$= \frac{1}{2}(\mathbf{t} - \mathbf{\Phi}\mathbf{w})^T(\mathbf{t} - \mathbf{\Phi}\mathbf{w})$$

where $\mathbf{t} = (t_1, \ldots, t_N)^T$, and

$$\mathbf{\Phi} = \begin{bmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \ldots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \ldots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \ldots & \phi_{M-1}(\mathbf{x}_N) \end{bmatrix}$$

plug into (3.11) $\quad \ln p(\mathbf{t} \,|\, \mathbf{w}, \beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})$

Then find the stationary point

$$\ln p(\mathbf{t} \mid \mathbf{w}, \beta) = \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w})$$
$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta \frac{1}{2} (\mathbf{t} - \mathbf{\Phi w})^T (\mathbf{t} - \mathbf{\Phi w})$$

The gradient with respect to $\mathbf{w}$ is

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t} \mid \mathbf{w}, \beta) = \beta \mathbf{\Phi}^T (\mathbf{t} - \mathbf{\Phi w}).$$

Setting the gradient to zero gives

$$0 = \mathbf{\Phi}^T \mathbf{t} - \mathbf{\Phi}^T \mathbf{\Phi w},$$

The normal equation: $$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^{\mathrm{T}} \mathbf{t} \tag{3.15}$$

Obtaining and interpreting MLE results

$$w_0 = \bar{t} - \sum_{j=1}^{M-1} w_j \overline{\phi_j} \qquad \bar{t} = \frac{1}{N}\sum_{n=1}^{N} t_n, \qquad \overline{\phi_j} = \frac{1}{N}\sum_{n=1}^{N} \phi_j(\mathbf{x}_n).$$

Recall: $$\ln p(\mathbf{t} \mid \mathbf{w}, \beta) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta\frac{1}{2}(\mathbf{t} - \mathbf{\Phi w})^T(\mathbf{t} - \mathbf{\Phi w})$$

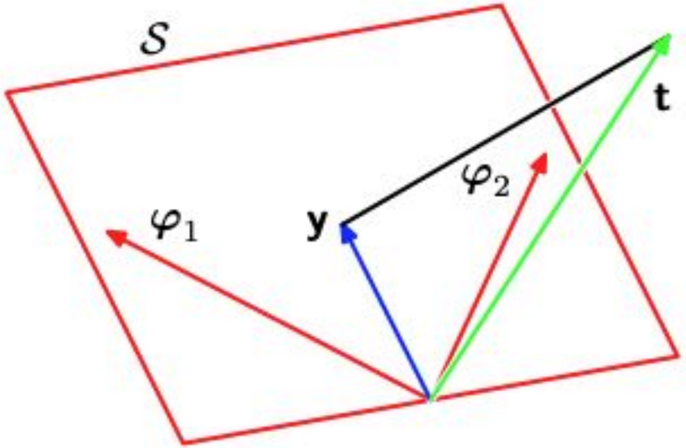
$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{t_n - \mathbf{w}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 \tag{3.21}$$

The normal equation:

$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t} \tag{3.15}$$

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$$

Figure 3.2 Geometrical interpretation of the least-squares solution, in an N-dimensional space whose axes are the values of $t_1, \ldots, t_N$. The least-squares regression function is obtained by finding the orthogonal projection of the data vector $\mathbf{t}$ onto the subspace spanned by the basis functions $\phi_j(\mathbf{x})$ in which each basis function is viewed as a vector $\boldsymbol{\varphi}_j$ of length N with elements $\phi_j(\mathbf{x}_n)$.



Numerical difficulties when $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ is ill-conditioned.

SVD or regularisation will help.

Cool geometric derivation of the normal equation
https://www.youtube.com/watch?v=PbyP3goun2Y

Sequential Learning - Stochastic Gradient Descent

- For large data sets, calculating the maximum likelihood parameters $\mathbf{w}_{ML}$ and β_{ML} may be costly.
- For online applications, never all data in memory.
- Use a sequential algorithms (online algorithm).
- If the error function is a sum over data points $E = \sum_n E_n$, then
  1. initialise $\mathbf{w}^{(0)}$ to some starting value
  2. update the parameter vector at iteration $\tau + 1$ by

Sum-of-squares error

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_n, \quad = \mathbf{w}^{(\tau)} + \eta \left(t_n - \mathbf{w}^{(\tau)T} \phi(\mathbf{x}_n) \right) \phi(\mathbf{x}_n)$$

where E_n is the error function after presenting the nth data set, and η is the learning rate.

What we did so far

Why linear models?

Input data, features, basis functions

Maximum likelihood and least squares

Geometric intuition, sequential update

Regularised least squares

Multiple outputs

Bias-variance decomposition

Regularised least squares

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \tag{3.24}$$

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}. \tag{3.27}$$

also (1.4)

$$\mathbf{w} = \left(\lambda\mathbf{I} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}. \tag{3.28}$$

Regularisation by q-norm

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q \tag{3.29}$$

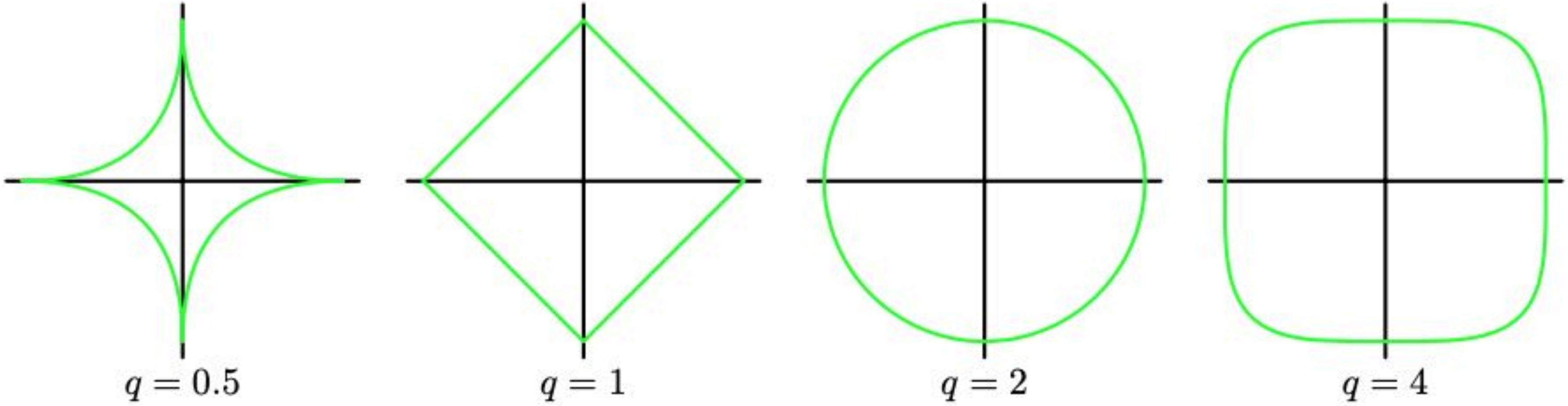


Figure 3.3 Contours of the regularization term in (3.29) for various values of the parameter q.

Lagrange multipliers (appendix E)

The first encounter in SML - we'll see it again in kernel methods.

objective function	minimize f(x)
equality constraint	subject to g(x)=0

$$\mathcal{L}(x, \lambda) = f(x) - \lambda g(x)$$

$\mathcal{L}$ = Lagrangian

λ = Lagrange multiplier

Figure E.1 A geometrical picture of the technique of Lagrange multipliers in which we seek to maximize a function $f(\mathbf{x})$, subject to the constraint $g(\mathbf{x}) = 0$. If $\mathbf{x}$ is D dimensional, the constraint $g(\mathbf{x}) = 0$ corresponds to a subspace of dimensionality $D - 1$, indicated by the red curve. The problem can be solved by optimizing the Lagrangian function $L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$.

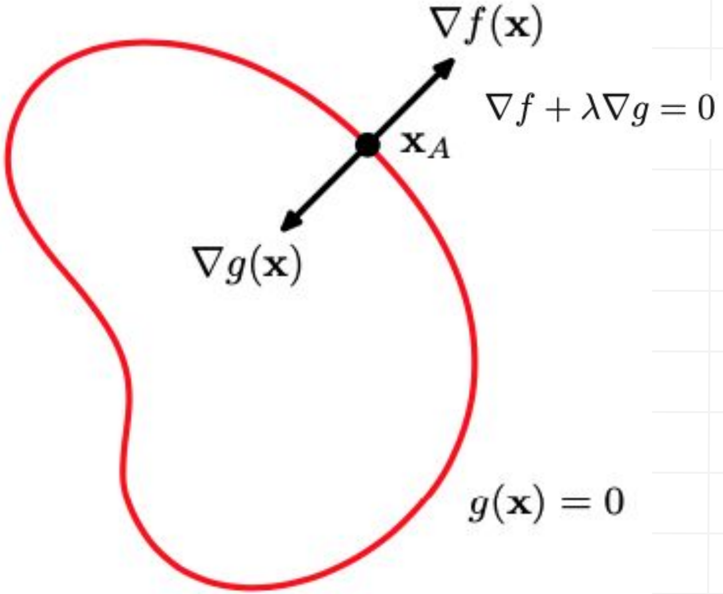


Figure E.2 A simple example of the use of Lagrange multipliers in which the aim is to maximize $f(x_1, x_2) = 1 - x_1^2 - x_2^2$ subject to the constraint $g(x_1, x_2) = 0$ where $g(x_1, x_2) = x_1 + x_2 - 1$. The circles show contours of the function $f(x_1, x_2)$, and the diagonal line shows the constraint surface $g(x_1, x_2) = 0$.

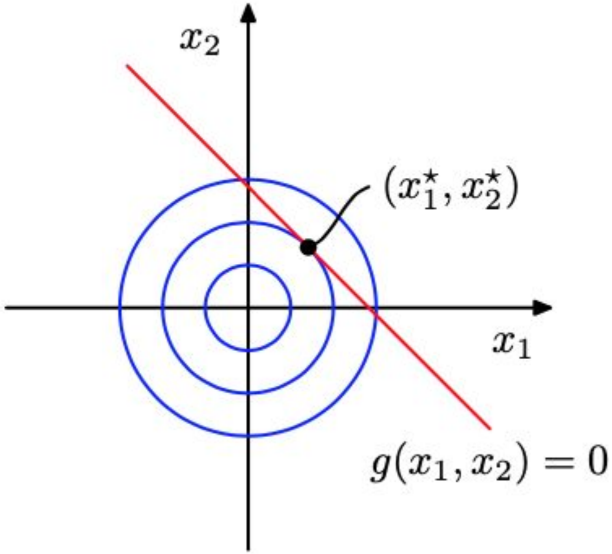
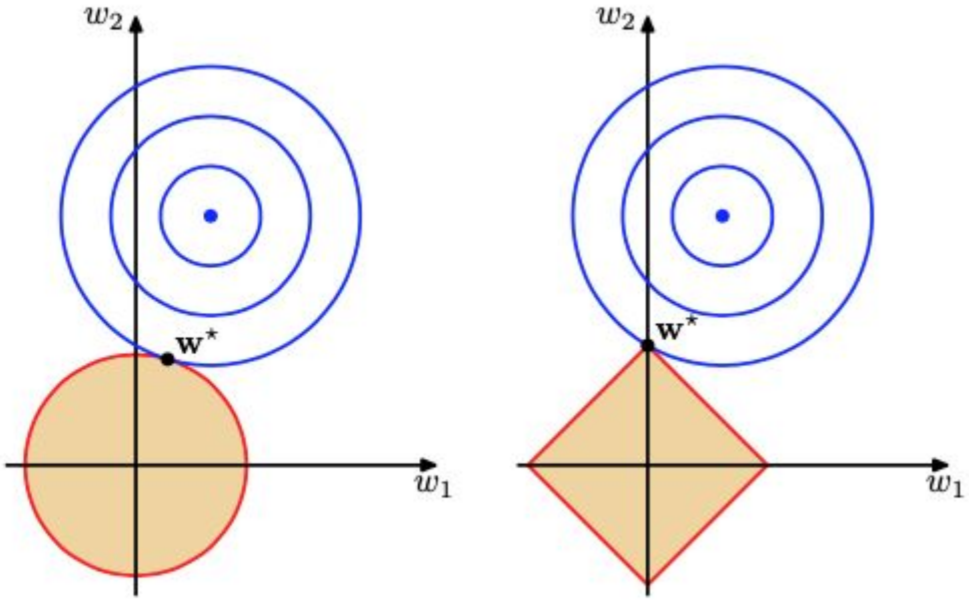


Figure 3.4 Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector $\mathbf{w}$ is denoted by $\mathbf{w}^{\star}$. The lasso gives a sparse solution in which $w_1^{\star} = 0$.



$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

Multiple regression outputs

See book

What we did so far

Why linear models?

Input data, features, basis functions

Maximum likelihood and least squares

Geometric intuition

Regularised least squares

Multiple outputs

Bias-variance decomposition

Bias-variance: the cartoon view

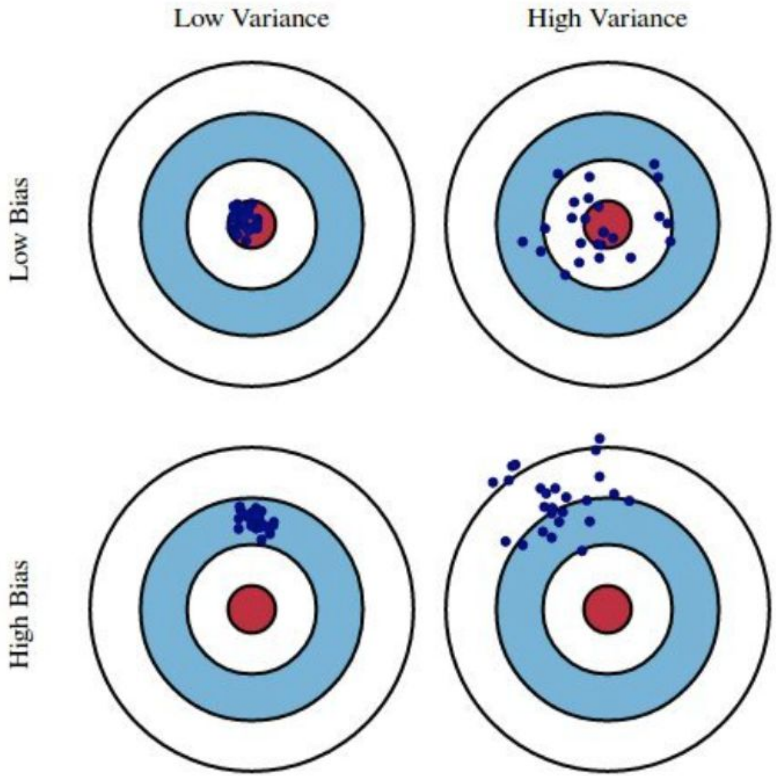


Fig. 1: Graphical illustration of bias and variance
From Understanding the Bias-Variance Tradeoff, by Scott Fortmann-Roe.

What are the sources of different dots in this picture?

What remains constant: learning target, model specification, learning algorithm.

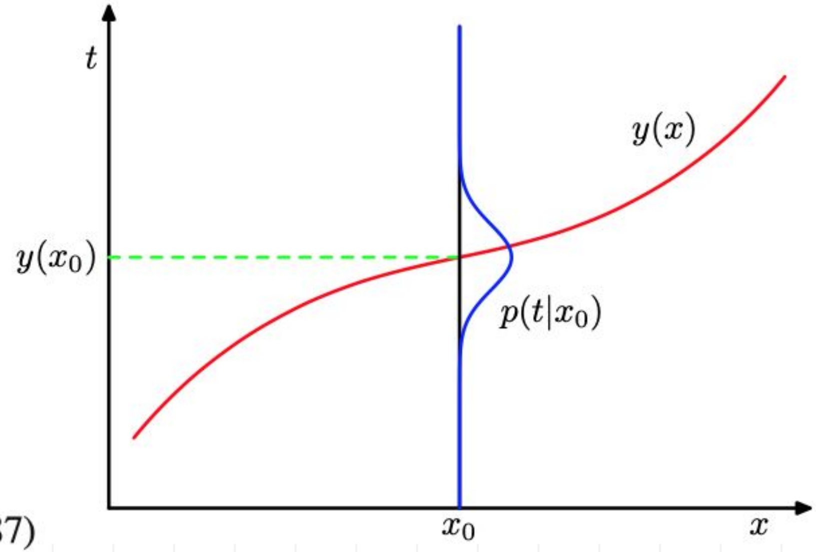
What changes:
Data (subsets), randomness in learning (including but are not limited to initialisations), ...

https://www.kdnuggets.com/2016/08/bias-variance-tradeoff-overview.html

Expected square loss

(See Sec 1.5.5)

Figure 1.28 The regression function $y(x)$, which minimizes the expected squared loss, is given by the mean of the conditional distribution $p(t|x)$.



$$\mathbf{y}(\mathbf{x}) = \mathbb{E}_t[\mathbf{t}|\mathbf{x}]$$

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x})\,\mathrm{d}t. \qquad (3.36)$$

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t. \qquad (1.87)$$

$$\begin{aligned}\{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\ &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2\end{aligned}$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x}. \qquad (1.90)$$

Generalising squared loss to Mikowski distances

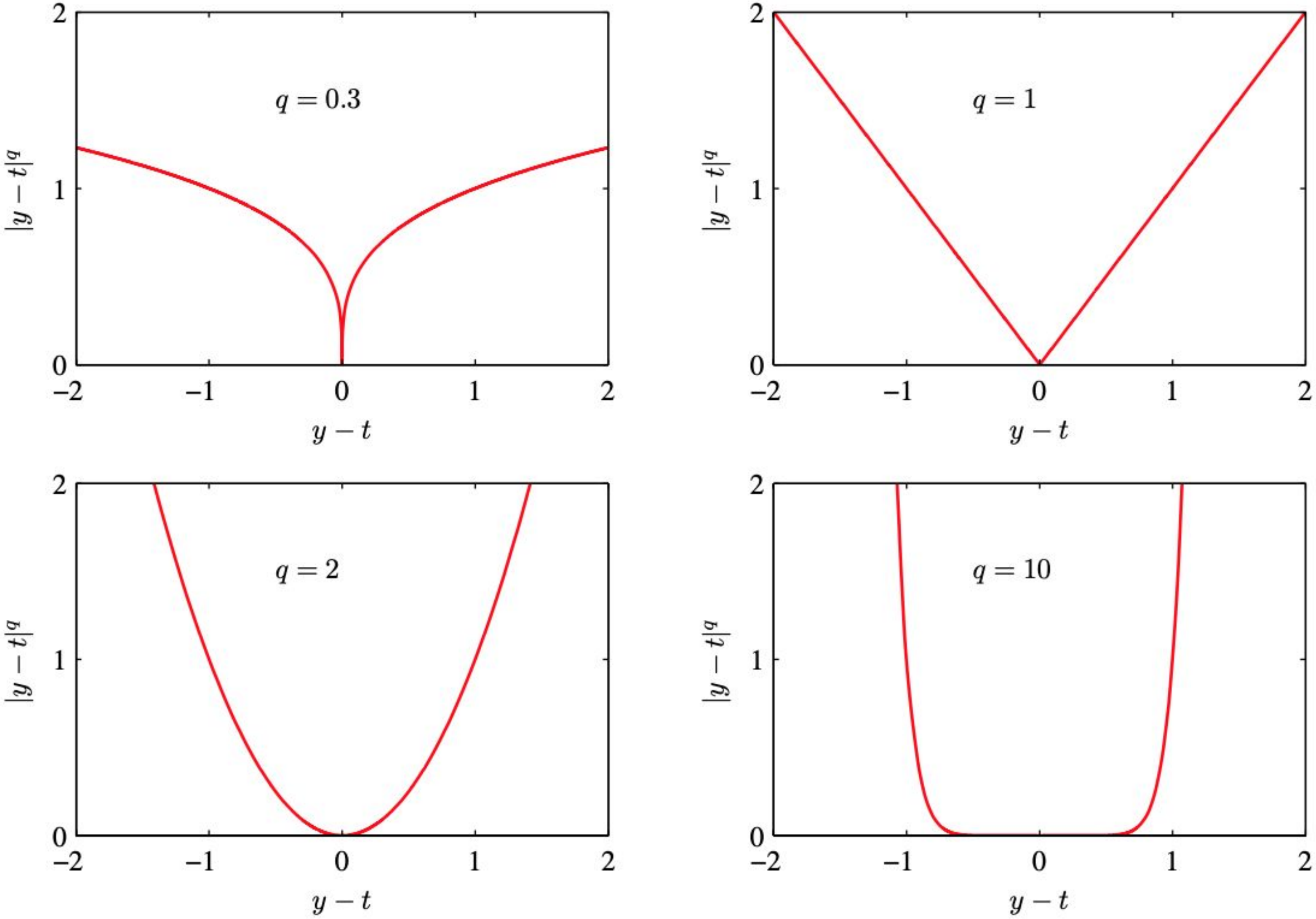


Figure 1.29 Plots of the quantity $L_q = |y - t|^q$ for various values of q.

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 \, p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \int \{\mathbb{E}[t|\mathbf{x}] - t\}^2 p(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{1.90}$$

$$h(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x}) \, \mathrm{d}t. \tag{3.36}$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 \, p(\mathbf{x}) \, \mathrm{d}\mathbf{x} + \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) \, \mathrm{d}\mathbf{x} \, \mathrm{d}t. \tag{3.37}$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t. \tag{3.37}$$

Now introduce data *D*

$$\{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2. \tag{3.38}$$

Q: how does one know to add+subtract E_D[y(x; D)] ?

$$= \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2$$

$$\begin{aligned} = \quad & \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\ & + 2\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}. \end{aligned} \tag{3.39}$$

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2\right] \\ & \quad = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})]\}^2\right]}_{\text{variance}}. \end{aligned} \tag{3.40}$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} + \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t. \tag{3.37}$$

$$\begin{aligned} &\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - h(\mathbf{x})\}^2\right] \\ &\quad = \underbrace{\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right]}_{\text{variance}}. \end{aligned} \tag{3.40}$$

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise} \tag{3.41}$$

where

$$(\text{bias})^2 = \int \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2 p(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{3.42}$$

$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{3.43}$$

$$\text{noise} = \int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t \tag{3.44}$$

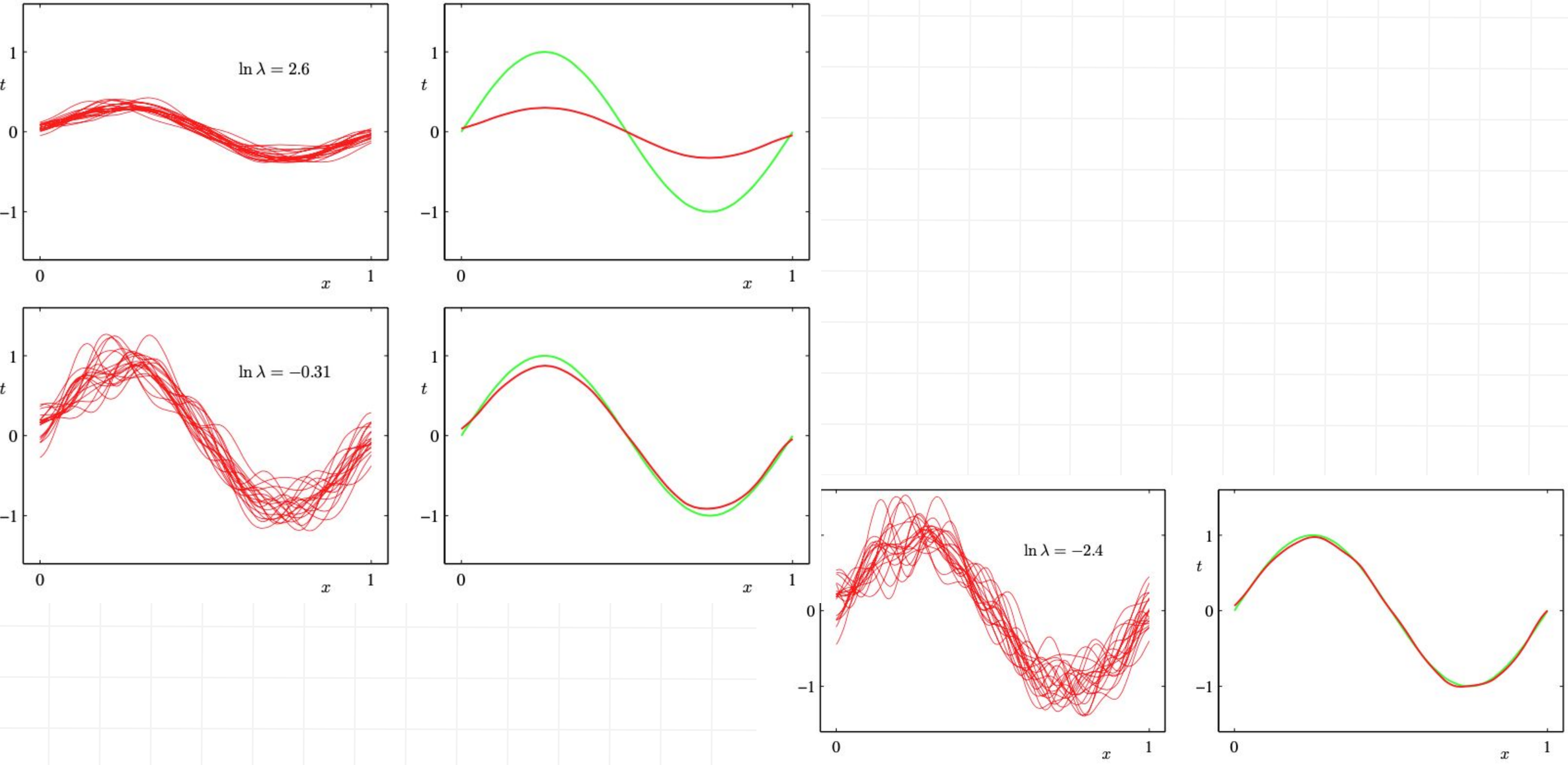
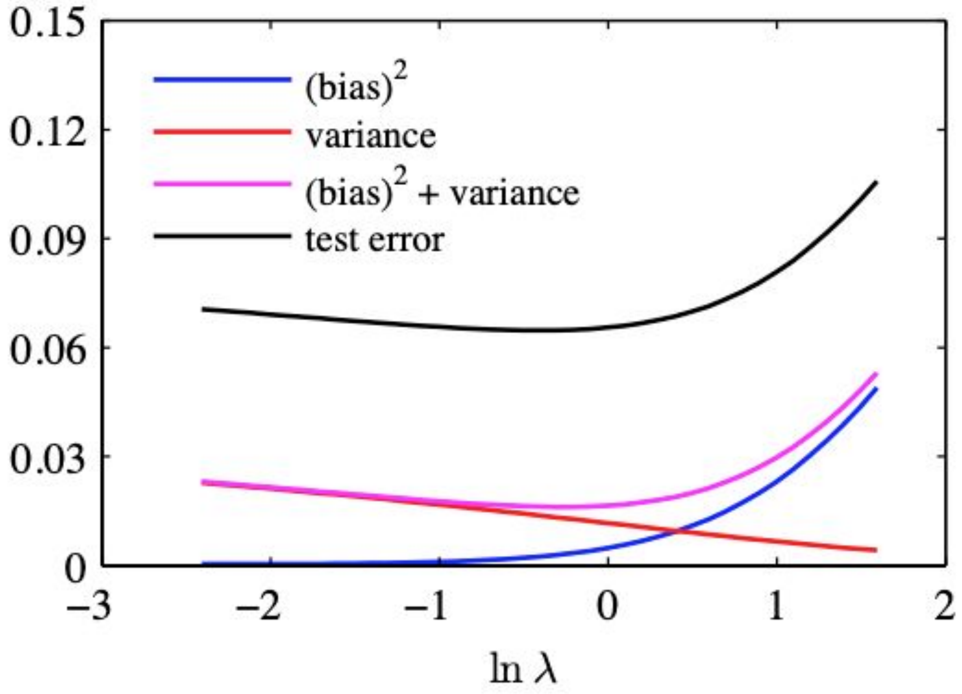


Figure 3.5 Illustration of the dependence of bias and variance on model complexity, governed by a regularization parameter λ, using the sinusoidal data set from Chapter 1. There are $L = 100$ data sets, each having $N = 25$ data points, and there are 24 Gaussian basis functions in the model so that the total number of parameters is $M = 25$ including the bias parameter. The left column shows the result of fitting the model to the data sets for various values of $\ln \lambda$ (for clarity, only 20 of the 100 fits are shown). The right column shows the corresponding average of the 100 fits (red) along with the sinusoidal function from which the data sets were generated (green).

- Dependence of bias and variance on the model complexity
- Squared bias, variance, their sum, and test data
- The minimum for $(\text{bias})^2 + \text{variance}$ occurs close to the value that gives the minimum error



Unbiased estimators

- You may have encountered *unbiased estimators*
- Why guarantee zero bias? To quote the pioneer of Bayesian inference, Edwin Jaynes, from his book *Probability Theory: The Logic of Science (2003)*:

Why do they do this? Why do orthodoxians put such exaggerated emphasis on bias? We suspect that the main reason is simply that they are caught in a psycho–semantic trap of their own making. When we call the quantity $(\langle\beta\rangle - \alpha)$ the "bias", that makes it sound like something awfully reprehensible, which we must get rid of at all costs. If it had been called instead the "component of error orthogonal to the variance", as suggested by the Pythagorean form of (17–2), it would have been clear to all that these two contributions to the error are on an equal footing; it is folly to decrease one at the expense of increasing the other. This is just the price one pays for choosing a technical terminology that carries an emotional load, implying value judgments; orthodoxy falls constantly into this tactical error.

The bias-variance decomposition

- Tradeoff between bias and variance
  - simple models have low variance and high bias
  - complex models have high variance and low bias
- The sum of bias and variance has a minimum at a certain model complexity.
- Expected loss $\mathbb{E}_{\mathcal{D}}[L]$ over all data sets $\mathcal{D}$

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

- The noise comes from the data, and can not be removed from the expected loss.
- To analyse the bias-variance decomposition : many data sets needed, which are not always available.

Sparse linear models and “explainable models”

Dawes, Robyn M.. “The robust beauty of improper linear models in decision making.” *American Psychologist* 34 (1979): 571-582.

Ustun, Berk and Cynthia Rudin. “Supersparse linear integer models for optimized medical scoring systems.” *Machine Learning* 102 (2015): 349-391.

Rudin, Cynthia. “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.” *Nature Machine Intelligence* 1 (2019): 206-215.

Cynthia Rudin, Duke U
2022 Squirrel Prize in AI for Humanity

Linear models for regression 1

Why linear models?

Input data, features, basis functions

Maximum likelihood and least squares

Geometric intuition

Regularised least squares

Multiple outputs

Bias-variance decomposition

The relation between MLE and least squares, Lagrange multipliers, multiple ways of looking at linear models.