# Linear Regression Models

Mingmin Chi

Fudan University, Shanghai, China

# Outline

# Supervised Learning

## Components for learning in common

- a set of variables –> inputs **x**, which are measured or preset
- one or more outputs (responses) $y$
- the goal is to use the inputs to predict the values of the outputs $\mathbf{x} -> \; y$

## Supervised learning

- given a set of data $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n$, where $\mathbf{x} \in \mathbb{R}^d, \; y \in \mathbb{R}$
- the prediction of a new sample **x** by $\mathcal{D}$, i.e., $y(\mathbf{x}|\mathcal{D})$ or $P(\mathbf{x}|\mathcal{D})$

# Function Approximation

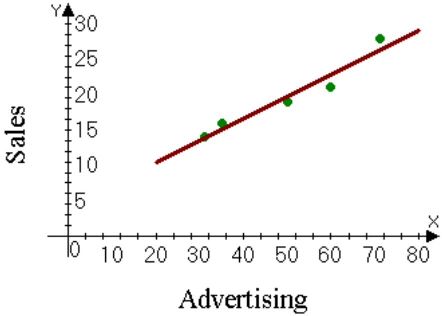
- If exists a mapping between inputs $\mathbf{x}$ and outputs $y$, the prediction can be obtained by *function approximation*, i.e., $y := f(\mathbf{x}, \mathbf{w})$
- What's the form of $f$?
- How to estimate $\mathbf{w}$?

# Probabilistic Distribution

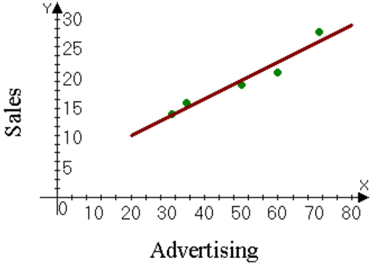
- uncertainty over the value of the target variable $t$ can be expressed by a probability distribution
- assume that given the value of $x$, the corresponding value of $t = p(t|x, \mathcal{D})$

# Regression

| Sales ($000,000s) ($y_i$) | Advertising ($000s) ($x_i$) |
|---|---|
| 28 | 71 |
| 14 | 31 |
| 19 | 50 |
| 21 | 60 |
| 16 | 35 |



# Regression (contd)

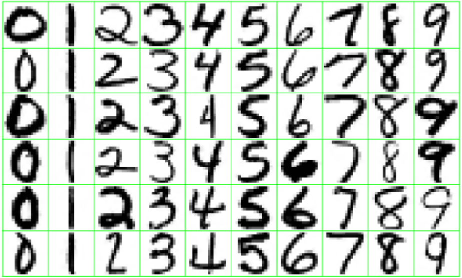



$$y = w_0 + w_1 x$$

The outputs $y$ is quantitative, the quantitative variables are *continuous* variable ⇒ regression when we predict quantitative outputs,

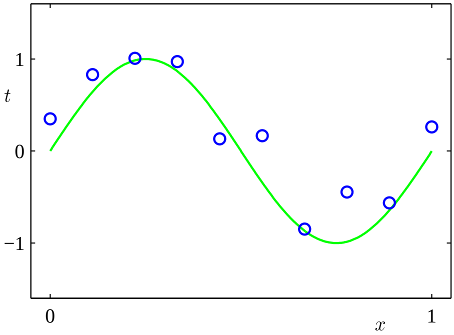
# Classification

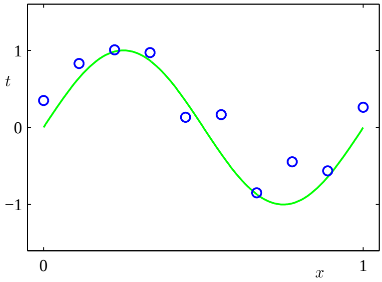
The outputs $y$ is qualitative, the qualitative variables are also referred to as *categorical* or *discrete* variable $\Rightarrow$, e.g., handwritten digit recognition, $C = \{0, 1, \cdots, 9\}$ classification when we predict qualitative outputs



## A simple regression problem

- observe a real-valued input variable $x$
- use this observation to predict the value of a real-valued target variable $t$
- consider synthetically generated data from the function $\sin(2\pi x)$ with random noise included in the target values






- given a training set comprising $N (N = 10)$ observations of $x$
- together with corresponding observations of the values of $t$
- the goal is to exploit this training set to make predictions of the value for new input variable

Difficulty: finite dataset; corruption with noise -> uncertainty to the appropriate value for $\hat{t}$

# Difficulty

- finite dataset
- corruption with noise

$\Rightarrow$ uncertainty to the appropriate value for $\hat{t}$

- probability theory
- decision theory

# Curve Fitting

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

where $M$ is the order of the polynomial and $x^j$ denotes $x$ raised to the power of $j$

# Curve Fitting

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

where $M$ is the order of the polynomial and $x^j$ denotes $x$ raised to the power of $j$

## Noted

- the polynomial function is a nonlinear function of $x$
- it is linear function of the coefficients $\mathbf{w}$

Functions, such as the polynomial, which are linear in the unknown parameters, are called linear models for regression

# Error Function

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

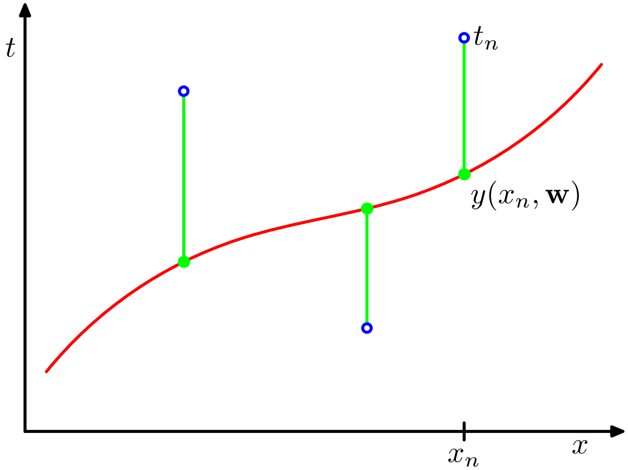
# Error Function

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

- the values of the coefficients can be determined by fitting the polynomial to the training data
- this can be done by minimizing an error function that measures the misfit between the function for any given value of **w** and the training set data points
- the sum of the squares of the errors (SSE) between the predictions $y(x_n, \mathbf{w})$ for each data point and the corresponding target values $t_n$:

$$\mathrm{Min} E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

# Geometrical Interpretation of SSE



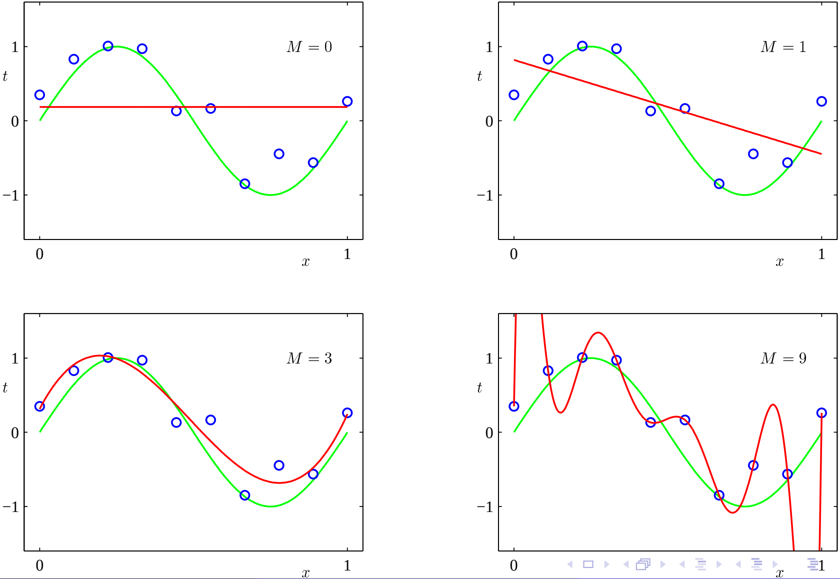
# Closed Form Solution of **w**

$$\mathrm{Min} E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

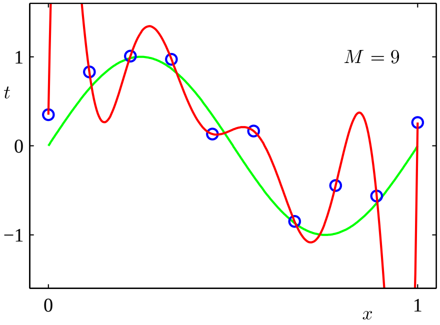
- the error function is a quadratic function of the coefficients **w**
- the derivatives w.r.t **w** will be linear in the elements of **w**
- the minimization of the error function has a unique solution denoted by $\mathbf{w}^*$

The resulting polynomial is given by the function $y(x, \mathbf{w}^*)$

# Choosing $M$: Model Selection



# Over-fitting

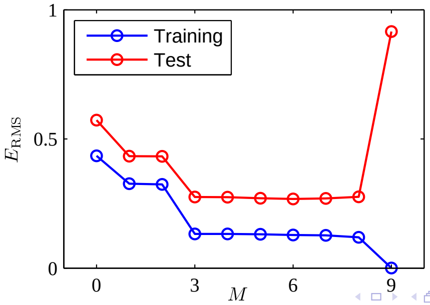



$E(\mathbf{w}^*) = 0$, but very poor representation of the function $\sin 2\pi x$, bad generalization

# RMS Errors

The goal of learning: to achieve good generalization by making accurate predictions for new data

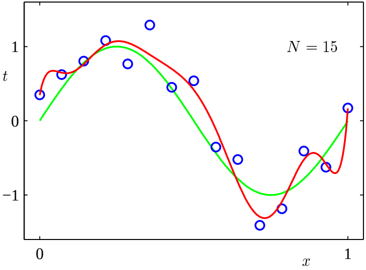
- training error
- test error
- root-mean-square (RMS) error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^*)/N}$
  - N for comparing different sizes of datasets in the same footing
  - the square root for measuring on the same scale as the target variable

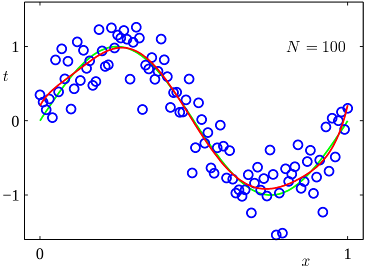


# Magnitude **w** with $M$

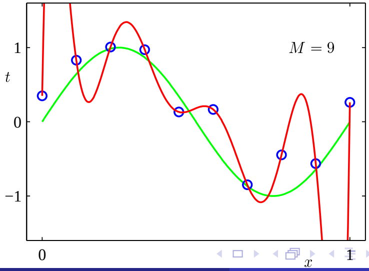
| | $M=0$ | $M=1$ | $M=3$ | $M=9$ |
|---|---|---|---|---|
| $w_0^\star$ | 0.19 | 0.82 | 0.31 | 0.35 |
| $w_1^\star$ | | -1.27 | 7.99 | 232.37 |
| $w_2^\star$ | | | -25.43 | -5321.83 |
| $w_3^\star$ | | | 17.37 | 48568.31 |
| $w_4^\star$ | | | | -231639.30 |
| $w_5^\star$ | | | | 640042.26 |
| $w_6^\star$ | | | | -1061800.52 |
| $w_7^\star$ | | | | 1042400.18 |
| $w_8^\star$ | | | | -557682.99 |
| $w_9^\star$ | | | | 125201.43 |

# More Training Data Points




$N = 10$



# Regularization

- Relatively complex and flexible models with limited training dataset
- e.g., curve fitting problem with $N = 10, M = 9$
- solution?

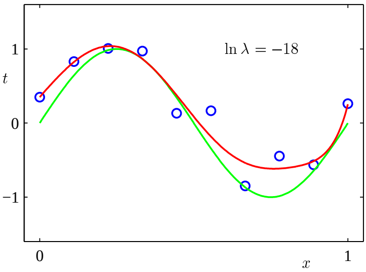
# Regularization

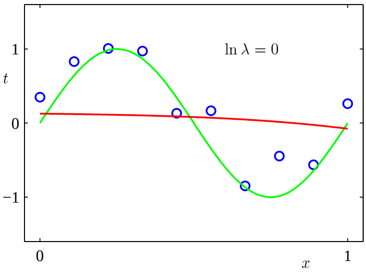
- Relatively complex and flexible models with limited training dataset
- e.g., curve fitting problem with $N = 10, M = 9$
- solution?

Regularization is used to control the over-fitting phenomenon, e.g.,

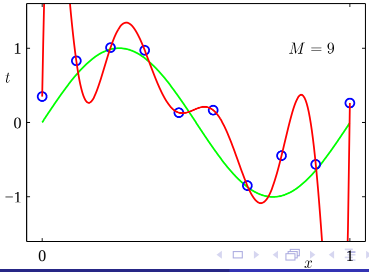
$$\tilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

# Regularization (contd)



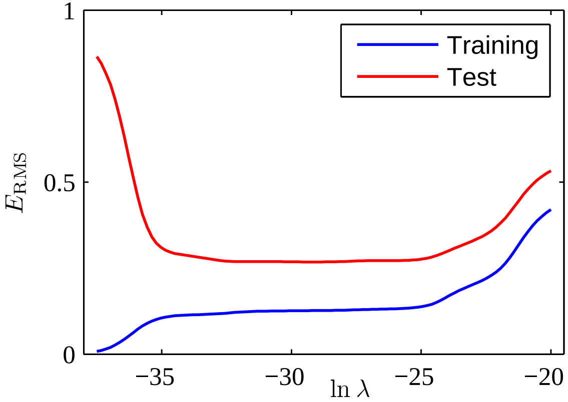

$\lambda = 0$, i.e., $\ln \lambda = -\infty$



# Magnitude **w** with Regularization

| | $\ln \lambda = -\infty$ | $\ln \lambda = -18$ | $\ln \lambda = 0$ |
|---|---|---|---|
| $w_0^\star$ | 0.35 | 0.35 | 0.13 |
| $w_1^\star$ | 232.37 | 4.74 | -0.05 |
| $w_2^\star$ | -5321.83 | -0.77 | -0.06 |
| $w_3^\star$ | 48568.31 | -31.97 | -0.05 |
| $w_4^\star$ | -231639.30 | -3.89 | -0.03 |
| $w_5^\star$ | 640042.26 | 55.28 | -0.02 |
| $w_6^\star$ | -1061800.52 | 41.32 | -0.01 |
| $w_7^\star$ | 1042400.18 | -45.95 | -0.00 |
| $w_8^\star$ | -557682.99 | -91.53 | 0.00 |
| $w_9^\star$ | 125201.43 | 72.68 | 0.01 |

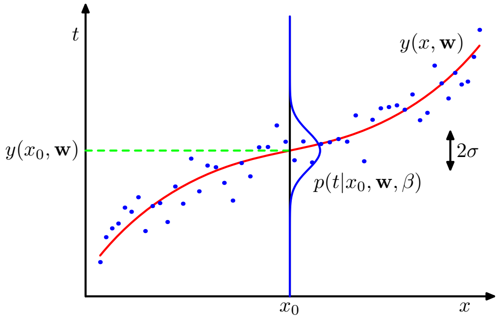
# RMS Errors with Regularization




$M = 9$, $\lambda$ controls the effective complexity of the model and determines the degree of over-fitting

- Assume that the target variable $t$ is given by a deterministic function $y(x, \mathbf{w})$ with additive Gaussian noise $\epsilon$
- Uncertainty over the value of the target variable $t$ can be expressed by a probability distribution
- Assume that $\epsilon \propto \mathcal{N}(t|0, \beta^{-1})$, then:

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



# Determination of **w**

$$p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n,\mathbf{w}),\beta^{-1})$$

$$\Rightarrow \ln p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

- $\mathbf{w}$ can be determined by maximum likelihood, denoted by $\mathbf{w}_{\mathrm{ML}}$
- considering $\mathbf{w}$, $\beta$ is constant -> $\max \ln p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)$ equivalently to $\min \frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2$,

# Determination of $\mathbf{w}$

$$p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n,\mathbf{w}),\beta^{-1})$$

$$\Rightarrow \ln p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

- $\mathbf{w}$ can be determined by maximum likelihood, denoted by $\mathbf{w}_{\mathrm{ML}}$
- considering $\mathbf{w}$, $\beta$ is constant -> max $\ln p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)$ equivalently to min $\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2$, the sum-of-squares error function
- the sum-of-squares error function has arisen as a consequence of maximizing likelihood under the assumption of a Gaussian noise distribution

# Determination of $\beta$

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n | y(x_n, \mathbf{w}), \beta^{-1})$$

$$\Rightarrow \ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

$\beta$ can be determined by maximum likelihood

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2$$

# MAP

- With $\mathbf{w}_{\text{ML}}$ and $\beta_{\text{ML}}$, we have

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}_{\text{ML}}), \beta_{\text{ML}}^{-1})$$

- Assume that a prior distribution over the coefficients $\mathbf{w}$, e.g., Gaussian distribution of the form

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{1}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}\right\}$$

Using Bayesian theorem, the posterior distribution for $\mathbf{w}$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

# MAP (contd)

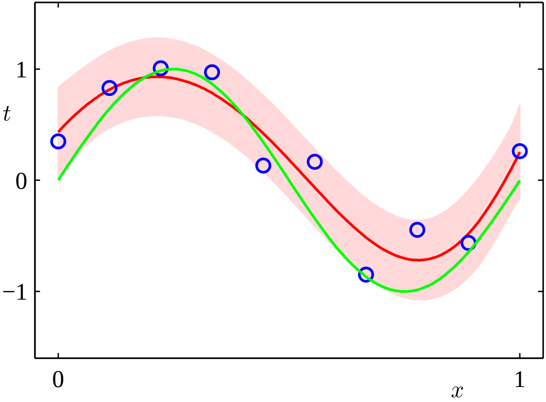
$$
\begin{aligned}
& p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) p(\mathbf{w}|\beta) \\
& \Rightarrow \ln p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \cdots \\
& \propto - \left\{ \frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} \right\}
\end{aligned}
$$

$\Rightarrow$ maximizing the posterior distribution is equivalent to minimizing the regularized sum-of-squares error function, with a regularization parameter given by $\lambda = \alpha/\beta$

# Bayesian Curve Fitting

- In the curve fitting problem, we are given the training data **x** and **t**,
- with a new test point $x$, the goal is to predict the value of $t$, i.e., the predictive distribution $p(t|x, \mathbf{x}, \mathbf{t})$
- $\alpha$ and $\beta$ are fixed and known in advance

$$
\begin{aligned}
p(t|x, \mathbf{x}, \mathbf{t}) &= \int p(t|x, \mathbf{w}) p(\mathbf{w}|\mathbf{x}, \mathbf{t}) d\mathbf{w} \\
&\propto \mathcal{N}(t|m(x), s^2(x))
\end{aligned}
$$


1
0
−1
t
0
1
x

# Regression Function

- Suppose that the decision stage consists of choosing a specific estimate $y(x)$ of the values of $t$ for each input $x$ and we incur a loss $\mathcal{L}(t, y(x))$:

$$\mathcal{E}[\mathcal{L}] = \int \int \mathcal{L}(t, y(x)) p(x, t) dx dt$$

# Regression Function

- Suppose that the decision stage consists of choosing a specific estimate $y(x)$ of the values of $t$ for each input $x$ and we incur a loss $\mathcal{L}(t, y(x))$:

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

- Our goal is to choose $y(x)$ so as to minimize $\mathcal{E}[\mathcal{L}]$
- If assume a completely flexible function $y(x)$, we can have

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

- Solving for $y(x)$ using the sum and product rules of probability, we obtain

$$y(x) = \frac{\int tp(x,t)dt}{p(x)} =$$

# Regression Function

- Suppose that the decision stage consists of choosing a specific estimate $y(x)$ of the values of $t$ for each input $x$ and we incur a loss $\mathcal{L}(t, y(x))$:

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

- Our goal is to choose $y(x)$ so as to minimize $\mathcal{E}[\mathcal{L}]$
- If assume a completely flexible function $y(x)$, we can have

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

- Solving for $y(x)$ using the sum and product rules of probability, we obtain

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

# Regression Function

- Suppose that the decision stage consists of choosing a specific estimate $y(x)$ of the values of $t$ for each input $x$ and we incur a loss $\mathcal{L}(t, y(x))$:

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

- Our goal is to choose $y(x)$ so as to minimize $\mathcal{E}[\mathcal{L}]$
- If assume a completely flexible function $y(x)$, we can have

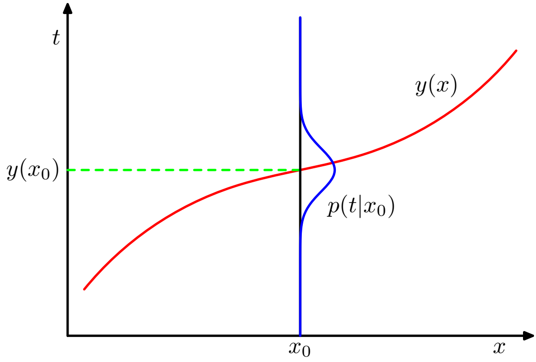
$$\frac{\partial \mathcal{E}[\mathcal{L}]}{\partial y(x)} = 2\int (y(x) - t)p(x,t)dt = 0$$

- Solving for $y(x)$ using the sum and product rules of probability, we obtain

$$y(x) = \frac{\int tp(x,t)dt}{p(x)} = \int tp(t|x)dt = \mathcal{E}_t[t|x]$$

# Regression Function (contd)

$y(x) = \int tp(t|x)dt = \mathcal{E}_t[t|x]$ is known as the regression function




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

# Three Approaches for Regression Problems

$y(x) = \int tp(x,t)dt = \mathcal{E}_t[t|x]$

- $p(x,t) \rightarrow^{p(x)} p(t|x) \rightarrow \int tp(x,t)dt$
- $p(t|x) \rightarrow \int tp(x,t)dt$
- Find a regression function $y(x)$ directly from the training data

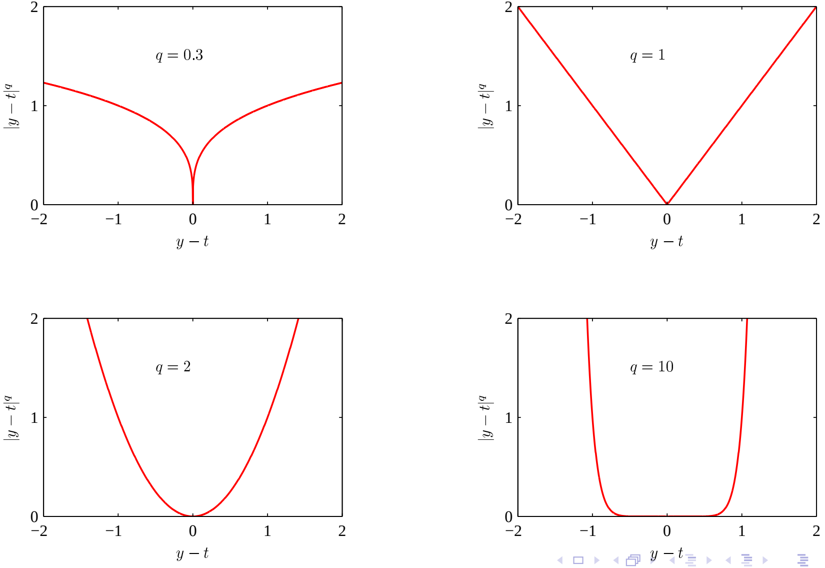
# Minkowski Loss

One simple generalization of the squared loss, called the Minkowski loss, whose expectation is given by

$$\mathcal{E}[\mathcal{L}_{\mathrm{q}}] = \int\int |y(x) - t|^q p(x, t) dxdt$$

- $q = 2$: the expected squared loss

# Minkowski Loss (contd)



# Linear Regression

The simplest linear model for regression is one that involves a linear combination of the input variables

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

where $\mathbf{x} = (x_1, \cdots, x_D)^\top$.

- This is often simply known as linear regression
- A linear function of the parameters $w_0, \cdots, \mathbf{w}_D$
- A linear function of the input variables $x_i$

# Basis Functions

- Limitation of the linear regression
- An extension by considering linear combinations of fixed nonlinear functions of the input variables:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

where $\phi_j(\mathbf{x})$ are know as basis function, e.g., in polynomial curve fitting, $\phi_j(\mathbf{x}) = x^j$

- $w_0$ is called a bias parameter. For convenience,

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x})$$

where $\mathbf{w} = (w_0, \cdots, w_{M-1})^\top$ and $\Phi = (\phi_0, \cdots, \phi_{M-1})^\top$

# Linear Regression: Revisit

- The simplest linear regression model

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

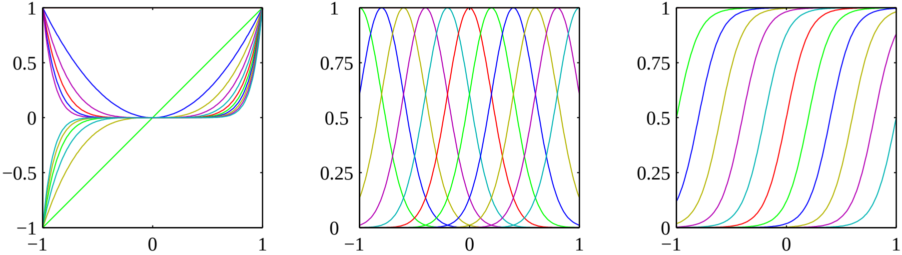
- By using nonlinear basis functions,

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

# Basis Functions (contd)

- Polynomial curve fitting, $\phi_j(x) = x^j$
- Gaussian basis functions, $\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$
- Sigmoidal basis functions, $\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$, where $\sigma(a)$ is the logistic sigmoid function defined by $\sigma(a) = \frac{1}{1+\exp(-a)}$

# Basis Functions (contd)



# Maximum Likelihood

- Assume that the target variable $t$ is given by a deterministic function $y(\mathbf{x}, \mathbf{w})$ with additive Gaussian noise so that

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

- $\epsilon = \mathcal{N}(0, \beta^{-1})$, thus we have

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

- Recall that

$$\mathcal{E}_{\mathbf{t}}[\mathbf{t}|\mathbf{x}] = \int \mathbf{t}p(\mathbf{t}|\mathbf{x})dt = y(\mathbf{x}, \mathbf{w})$$

- the likelihood function of the adjustable parameters $\mathbf{w}$ and $\beta$:

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

# Determination of $\mathbf{w}_{\mathrm{ML}}$

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^\top \Phi(\mathbf{x}_n), \beta^{-1})$$

$$\Rightarrow \ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^\top \Phi(\mathbf{x}_n), \beta^{-1})$$

$$= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w})$$

where $E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - \mathbf{w}^\top \Phi(\mathbf{x}_n)\}^2$.

We can use maximum likelihood to determine $\mathbf{w}$ and $\beta$:

$$\nabla \ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \sum_{n=1}^{N} \{t_n - \mathbf{w}^\top \Phi(\mathbf{x}_n)\} \Phi(\mathbf{x}_n)^\top$$

# Determination of $\mathbf{w}_{\mathrm{ML}}$ and $\beta_{\mathrm{ML}}$

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \sum_{n=1}^{N} \{t_n - \mathbf{w}^\top \Phi(\mathbf{x}_n)\} \Phi(\mathbf{x}_n)^\top$$

$$\Rightarrow \mathbf{w}_{\mathrm{ML}} = (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{t}$$

$$\nabla_{\beta} \ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) \Rightarrow \frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{\mathbf{t}_n - \mathbf{w}^\top \Phi(\mathbf{x}_n)\}^2$$

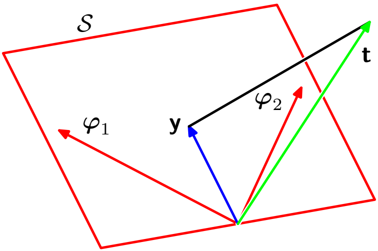
# Pseudo-Inverse of A Matrix

$\Phi \in \mathbb{R}^{N\times M}$

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

- Moore-Penrose pseudo-inverse of the matrix $\Phi$:$\Phi^{\dagger} \equiv (\Phi^{\top}\Phi)^{-1}\Phi^{\top}$

# Geometry of Least Squares




- $M < N$, $\mathcal{S} = \text{span}(\varphi_1, \cdots, \varphi_{M-1})$
- $\mathbf{y}$ can live anywhere in the $M-$dimensional subspace
- $E_D(\mathbf{w}) = ||\mathbf{y} - \mathbf{t}||^2$
- the least-squares solution for $\mathbf{w}$ corresponds to that choice of $\mathbf{y}$ that lies in subspace $\mathcal{S}$ and that is closest to $\mathbf{t}$
- the solution corresponds to the orthogonal projection of $\mathbf{t}$ onto the subspace $\mathcal{S}$

Numerical difficulty when $\Phi^\top\Phi$ is close to singular,e.g., when two or more of the basis vectors $\varphi_j$ are co-linear, or nearly so

**Possible solutions**

- singular value decomposition
- regularization

# Regularized Least Squares

- To control over-fitting, total error function takes the form

$$\tilde{E}(\mathbf{w}) = E_D(\mathbf{w}) + \lambda E_{\mathbf{w}}(\mathbf{w})$$

- one of the simplest forms of regularizer is given by

$$E_{\mathbf{w}}(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbf{w}$$

- if the sum-of-squares error function is taken, then total error functions

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top \Phi(\mathbf{x}_n)\}^2 + \frac{1}{2}\mathbf{w}^\top \mathbf{w}$$

- the close-formed solution for $\mathbf{w}$ is

$$\mathbf{w} = (\lambda \mathbf{I} + \Phi^\top \Phi)^{-1}\Phi^\top \mathbf{t}$$

# Regularized Least Squares

- To control over-fitting, total error function takes the form

$$\tilde{E}(\mathbf{w}) = E_D(\mathbf{w}) + \lambda E_{\mathbf{w}}(\mathbf{w})$$

- one of the simplest forms of regularizer is given by

$$E_{\mathbf{w}}(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \mathbf{w}$$

- if the sum-of-squares error function is taken, then total error functions

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top \Phi(\mathbf{x}_n)\}^2 + \frac{1}{2}\mathbf{w}^\top \mathbf{w}$$

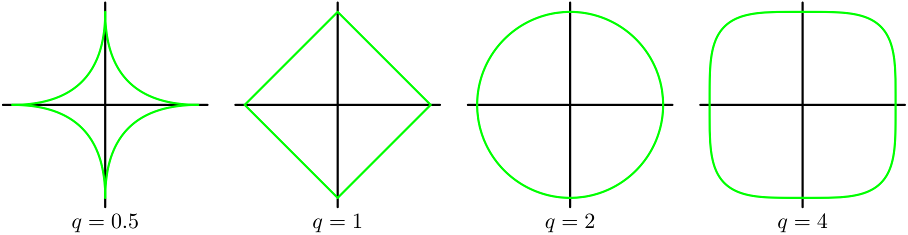
- the close-formed solution for **w** is

$$\mathbf{w} = (\lambda \mathbf{I} + \Phi^\top \Phi)^{-1}\Phi^\top \mathbf{t}$$

# Regularizers

A more general regularizer is sometimes used

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top \Phi(\mathbf{x}_n)\}^2 + \frac{1}{2}\sum_{j=1}^{M}|\mathbf{w}|^q$$



# Over-fitting Problem

## Linear models for regression

Fixing the form and the number of basis functions

- Over-fitting for complex models trained by datasets of limited size, e.g., ML or least square
- Loss of flexibility of the model by limiting the number of basis function to avoid over-fitting
- How to determine $\lambda$ by the introduction of regularization terms to control over-fitting

Over-fitting for MLE but not in a Bayesian setting when we marginalize over parameters

# Expected Squared Loss: Revisited

- Given the conditional distribution $p(t|\mathbf{x})$
- Optimal prediction

$$h(\mathbf{x}) = \mathcal{E}[t|\mathbf{x}] = \int tp(t|\mathbf{x})dt.$$

- Squared loss function:

$$\{y(\mathbf{x}) - t\}^2 = \{y(\mathbf{x}) - \mathcal{E}[t|\mathbf{x}] + \mathcal{E}[t|\mathbf{x}] - t\}^2$$
$$= \{y(\mathbf{x}) - \mathcal{E}[t|\mathbf{x}]\}^2 + \{\mathcal{E}[t|\mathbf{x}] - t\}^2 + 2\{y(\mathbf{x}) - \mathcal{E}[t|\mathbf{x}]\}\{\mathcal{E}[t|\mathbf{x}] - t\}$$

- Expected squared loss function:

$$\mathcal{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x}) d\mathbf{x} + \underbrace{\int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt}_{\text{independent of } y(\mathbf{x})\text{; intrinsic noise on the data}}$$

# Expected Squared Loss: Revisited

- Given the conditional distribution $p(t|\mathbf{x})$
- Optimal prediction

$$h(\mathbf{x}) = \mathcal{E}[t|\mathbf{x}] = \int tp(t|\mathbf{x})dt.$$

- Squared loss function:

$$\{y(\mathbf{x}) - t\}^2 = \{y(\mathbf{x}) - \mathcal{E}[t|\mathbf{x}] + \mathcal{E}[t|\mathbf{x}] - t\}^2$$
$$= \{y(\mathbf{x}) - \mathcal{E}[t|\mathbf{x}]\}^2 + \{\mathcal{E}[t|\mathbf{x}] - t\}^2 + 2\{y(\mathbf{x}) - \mathcal{E}[t|\mathbf{x}]\}\{\mathcal{E}[t|\mathbf{x}] - t\}$$

- Expected squared loss function:

$$\mathcal{E}[L] = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x}) d\mathbf{x} + \underbrace{\int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt}_{\text{independent of } y(\mathbf{x})\text{; intrinsic noise on the data}}$$

# Expected Squared Loss (contd)

- Modeling $h(\mathbf{x})$ using a parametric function $y(\mathbf{x}, \mathbf{w})$
- Uncertainty in the model from a Bayesian perspective being expressed by a posterior distribution over $\mathbf{w}$
- Estimation of $\mathbf{w}$ based on the dataset $\mathcal{D}$ in a frequentist treatment
- Obtaining different prediction functions $y(\mathbf{x}, \mathcal{D})$ based on different datasets $\Longrightarrow$ different values of the squared loss
- The performance of a particular learning algorithm is assessed by taking the average over this ensemble of datasets

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

- Dependent on the particular dataset $\mathcal{D}$
- Taking its average over the ensemble of datasets:

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

- the expectation of the expression wrt $\mathcal{D}$

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

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

where

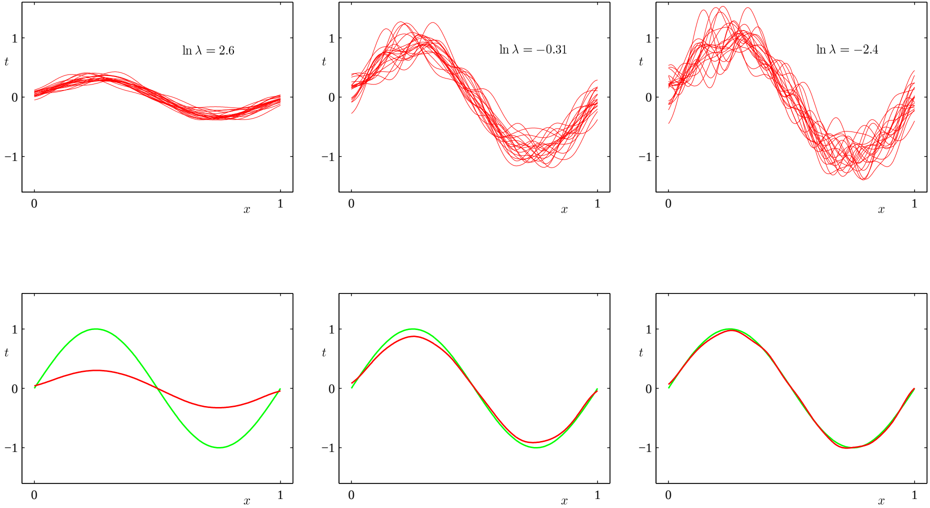
$$(\text{bias})^2 = \{\mathcal{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})] - h(\mathbf{x})\}^2$$

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

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

Our goal is to minimize the expected loss

- trade-off between bias and variance
- flexible models having low bias and high variance
- rigid models having high bias and low variance



$\ln \lambda = 2.6$
$\ln \lambda = -0.31$
$\ln \lambda = -2.4$
$t$
$x$
1
0
−1

Result of averaging many solutions for the complex model is a very good fit to the regression function

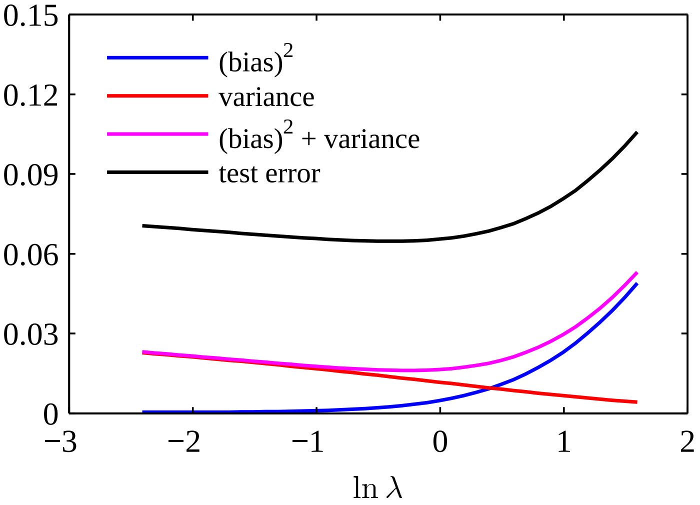
- averaging might be a beneficial procedure
- the average prediction is estimated from

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

and the integrated squared bias and integrated variance

$$(\text{bias})^2 = \frac{1}{N}\sum_{n=1}^{N}\{\bar{y}(\mathbf{x}) - h(\mathbf{x})\}^2$$

$$\text{variance} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{n=1}^{L} y^{(l)}(\mathbf{x}) - \bar{y}(\mathbf{x})\}^2$$


0.15
0.12
0.09
0.06
0.03
0
−3
−2
−1
0
1
2
(bias)$^2$
variance
(bias)$^2$ + variance
test error
ln $\lambda$