Artificial Intelligence

Math Notation and Basic Concept

Donghui Wang
AI Institute@ZJU
2015.03

Contents

- Linear algebra review and notation
- Example: Polynomial curve fitting
- Probability theory review and notation
- Curve fitting: probabilistic perspective

References:

1. *Zico Kolter. "Linear algebra review and reference", 2012.*
2. *Gene H. Golub, Charles F. Van Loan. "Matrix Computations",2009.*
3. *Christopher M. Bishop. "Pattern Recognition and Machine Learning",2006.*
4. *http://cs229.stanford.edu/materials.html*
5. *The Matrix Cookbook. http://matrixcookbook.com*

Linear algebra review and notation

Vector and Matrix

We use the following notation:

- By $A \in \mathbb{R}^{m \times n}$ we denote a matrix with m rows and n columns, where the entries of A are real numbers.
- By $x \in \mathbb{R}^n$, we denote a vector with n entries. By convention, an n-dimensional vector is often thought of as a matrix with n rows and 1 column, known as a ***column vector***. If we want to explicitly represent a ***row vector*** — a matrix with 1 row and n columns — we typically write x^T (here x^T denotes the transpose of x, which we will define shortly).
- The ith element of a vector x is denoted x_i:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Vector and Matrix

- We use the notation a_{ij} (or A_{ij}, $A_{i,j}$, etc) to denote the entry of A in the ith row and jth column:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

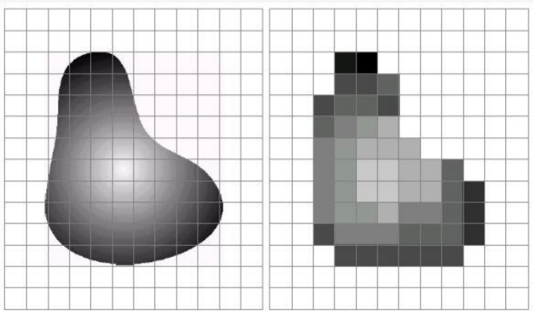
- We denote the jth column of A by a_j or $A_{:,j}$:

$$A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix}.$$

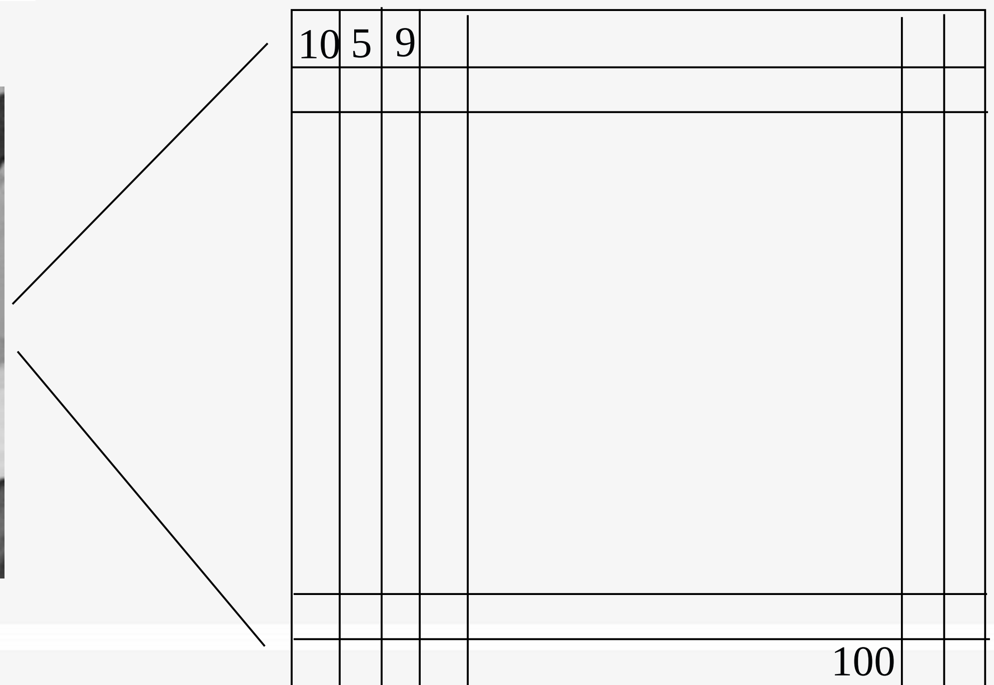
- We denote the ith row of A by a_i^T or $A_{i,:}$:

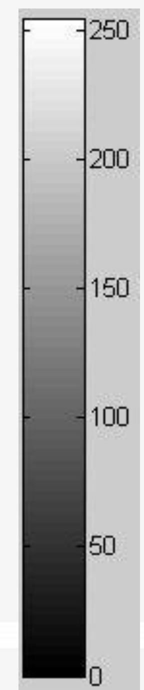
$$A = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix}.$$

Vector and Matrix



```
I = imread('xxx.jpg');
A = rgb2gray(I);
figure, imshow(I);
figure, imshow(A);
b = A( : );       %matrix to vector
```





Matrix multiplication

- Matrix-matrix products:

The product of two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$ is the matrix

$$C = AB \in \mathbb{R}^{m\times p},$$

where

$$C_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}.$$

```
A = rand(2,3);
B = rand(3,4);
C = A*B;
```

```
A = rand(2,3);
B = rand(2,3);
C = A.*B;
```

$$C = AB = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & \cdots & a_1^T b_p \\ a_2^T b_1 & a_2^T b_2 & \cdots & a_2^T b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^T b_1 & a_m^T b_2 & \cdots & a_m^T b_p \end{bmatrix} = \begin{bmatrix} — & a_1^T B & — \\ — & a_2^T B & — \\ & \vdots & \\ — & a_m^T B & — \end{bmatrix}$$

$$C = AB = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} — & b_1^T & — \\ — & b_2^T & — \\ & \vdots & \\ — & b_n^T & — \end{bmatrix} = \sum_{i=1}^{n} a_i b_i^T = \begin{bmatrix} | & | & & | \\ Ab_1 & Ab_2 & \cdots & Ab_p \\ | & | & & | \end{bmatrix}$$

Matrix multiplication

- Vector-vector products:

Given two vectors $x, y \in \mathbb{R}^n$, the quantity $x^T y$, sometimes called the ***inner product*** or ***dot product*** of the vectors, is a real number given by

$$x^T y \in \mathbb{R} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

Observe that inner products are really just special case of matrix multiplication. Note that it is always the case that $x^T y = y^T x$.

Given vectors $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ (not necessarily of the same size), $xy^T \in \mathbb{R}^{m \times n}$ is called the ***outer product*** of the vectors. It is a matrix whose entries are given by $(xy^T)_{ij} = x_i y_j$, i.e.,

$$xy^T \in \mathbb{R}^{m\times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}.$$

Matrix multiplication

```
A = rand(2,3);
x = rand(3,1);
y = A*x;
```

- Matrix-vector products:

Given a matrix $A \in \mathbb{R}^{m\times n}$ and a vector $x \in \mathbb{R}^n$, their product is a vector $y = Ax \in \mathbb{R}^m$. There are a couple ways of looking at matrix-vector multiplication, and we will look at each of them in turn.

If we write A by rows, then we can express Ax as,

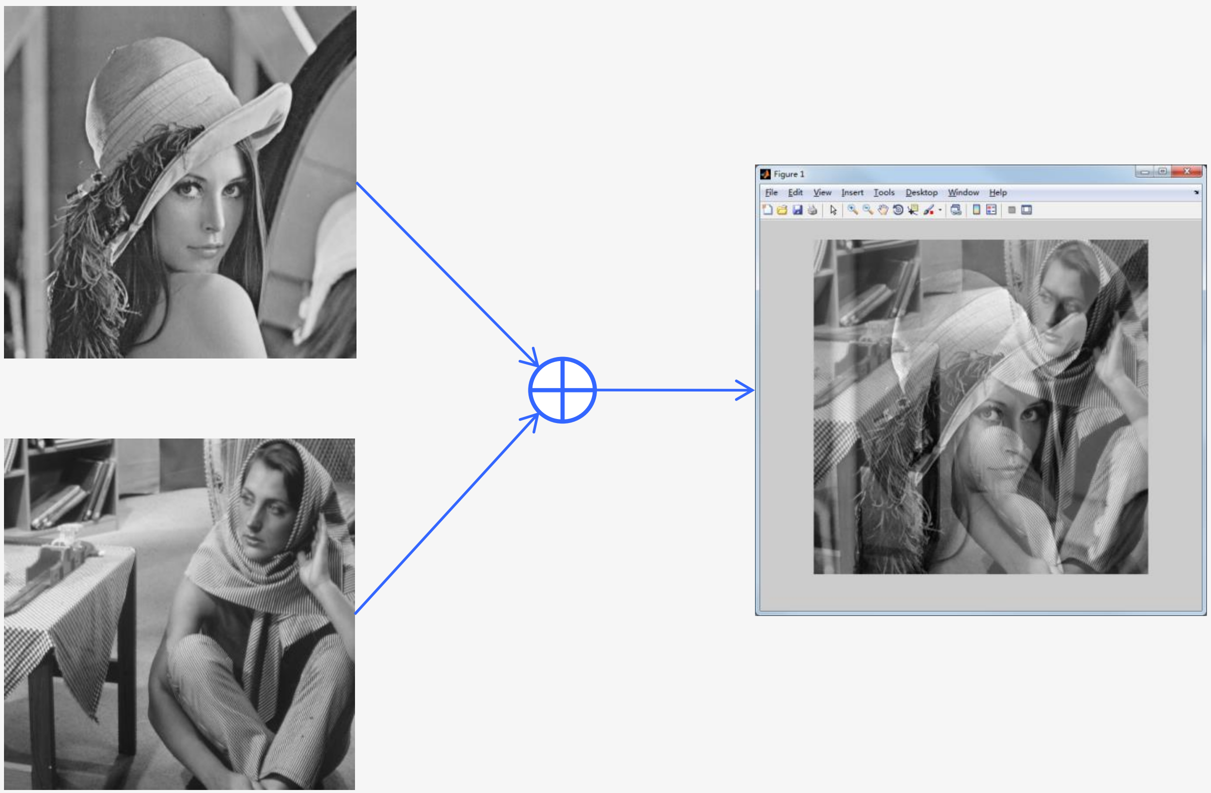
$$y = Ax = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{bmatrix}.$$

$$y = Ax = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} a_1 \end{bmatrix} x_1 + \begin{bmatrix} a_2 \end{bmatrix} x_2 + \ldots + \begin{bmatrix} a_n \end{bmatrix} x_n \ .$$

In other words, y is a ***linear combination*** *of the* *columns* *of* A, where the coefficients of the linear combination are given by the entries of x.

Matrix multiplication

- Matrix-vector products:




```
I1 = imread('lena.png');
a1 = double(I1(:));
I2 = imread('barbara.png');
a2 = double(I2(:));
A = [a1 a2];
x = rand(2, 1);
y = A*x;
I3 = reshape(y, size(I2));
figure, imshow( I3, [ ] );
```

Operations and properties

- The identity matrix: $I \in \mathbb{R}^{n \times n}$ $\quad I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$

- Diagonal matrix: $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ $\quad D_{ij} = \begin{cases} d_i & i = j \\ 0 & i \neq j \end{cases}$

- The transpose: $(A^T)^T = A \qquad (AB)^T = B^T A^T$

$$(A + B)^T = A^T + B^T$$

- Symmetric matrices: $A = A^T$

- The inverse: $A^{-1}A = I = AA^{-1} \qquad (A^{-1})^{-1} = A$

$$(AB)^{-1} = B^{-1}A^{-1} \qquad (A^{-1})^T = (A^T)^{-1}$$

- The determinant:

$$|A| \text{ or } \det A \quad |A| = |A^T| \quad |AB| = |A||B| \quad |A^{-1}| = 1/|A|$$

```
A = eye(2,3,'int8');

A = ones(2,3,'int8');
A = zeros(2,3,'int8');

d = rand(3,1);
A = diag(d);

y = x';
B = A';

Y = inv(X);
t = det(X);
```

Operations and properties

- The trace of a square matrix *A*:

$$\mathrm{tr}A = \sum_{i=1}^{n} A_{ii}$$

```
X = rand(4, 4)
t = trace(X)
v = diag(X)
s = sum(v)

d = rand(3,1)
A = diag(d)
t = trace(A)
s = sum(d)
```

- For $A \in \mathbb{R}^{n\times n}$, $\mathrm{tr}A = \mathrm{tr}A^T$.
- For $A, B \in \mathbb{R}^{n\times n}$, $\mathrm{tr}(A+B) = \mathrm{tr}A + \mathrm{tr}B$.
- For $A \in \mathbb{R}^{n\times n}$, $t \in \mathbb{R}$, $\mathrm{tr}(tA) = t\,\mathrm{tr}A$.
- For A, B such that AB is square, $\mathrm{tr}AB = \mathrm{tr}BA$.
- For A, B, C such that ABC is square, $\mathrm{tr}ABC = \mathrm{tr}BCA = \mathrm{tr}CAB$, and so on for the product of more matrices.

Operations and properties

- Norms:

A ***norm*** of a vector $\|x\|$ is informally a measure of the "length" of the vector. For example, we have the commonly-used Euclidean or ℓ_2 norm,

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Note that $\|x\|_2^2 = x^T x$.

More formally, a norm is any function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ that satisfies 4 properties:

1. For all $x \in \mathbb{R}^n$, $f(x) \geq 0$ (non-negativity).
2. $f(x) = 0$ if and only if $x = 0$ (definiteness).
3. For all $x \in \mathbb{R}^n$, $t \in \mathbb{R}$, $f(tx) = |t| f(x)$ (homogeneity).
4. For all $x, y \in \mathbb{R}^n$, $f(x + y) \leq f(x) + f(y)$ (triangle inequality).

Operations and properties

n = norm(x)
n = norm(x,p)

n = norm(A)
n = norm(A,p)

- Vector norms:
  - L_p norm
  - L_2 (*Euclidean*) norm
  - L_1 norm
  - L_∞ norm

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} \qquad \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i| \qquad \|x\|_\infty = \max_i |x_i|$$

- Matrix norms:
  - *Frobenius* norm (*Hilbert-Schmidt* norm):
  - Nuclear norm (trace norm)
  - Spectral norm

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2} = \sqrt{\mathrm{tr}(A^T A)}$$

$$\|A\|_* = \mathrm{trace}\left(\sqrt{A^* A}\right) = \sum_{i=1}^{\min\{m,n\}} \sigma_i.$$

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^* A)} = \sigma_{\max}(A)$$

Operations and properties

- Linear independence and rank:

A set of vectors $\{x_1, x_2, \ldots x_n\} \subset \mathbb{R}^m$ is said to be ***(linearly) independent*** if no vector can be represented as a linear combination of the remaining vectors.

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i$$

  – *The column (row) rank is referred to as the number of linearly independent columns (rows) of A, denoted as rank(A).*

- Orthogonal matrices and normalized matrices

  – Two vectors $x, y \in \mathbb{R}^n$ are ***orthogonal*** if $x^T y = 0$.

  – A vector $x \in \mathbb{R}^n$ is ***normalized*** if $\|x\|_2 = 1$.

  · A square matrix $U \in \mathbb{R}^{n \times n}$ is ***orthogonal*** if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being ***orthonormal***).

Operations and properties

- Span, range and nullspace:

The ***span*** of a set of vectors $\{x_1, x_2, \ldots x_n\}$ is the set of all vectors that can be expressed as a linear combination of $\{x_1, \ldots, x_n\}$. That is,

$$\operatorname{span}(\{x_1, \ldots x_n\}) = \left\{ v : v = \sum_{i=1}^{n} \alpha_i x_i, \;\; \alpha_i \in \mathbb{R} \right\}.$$

The ***range*** (sometimes also called the columnspace) of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{R}(A)$, is the the span of the columns of A. In other words,

$$\mathcal{R}(A) = \{v \in \mathbb{R}^m : v = Ax, x \in \mathbb{R}^n\}.$$

The ***nullspace*** of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{N}(A)$ is the set of all vectors that equal 0 when multiplied by A, i.e.,

$$\mathcal{N}(A) = \{x \in \mathbb{R}^n : Ax = 0\}.$$

Operations and properties

- Quadratic forms

Given a square matrix $A \in \mathbb{R}^{n \times n}$ and a vector $x \in \mathbb{R}^n$, the scalar value $x^T A x$ is called a ***quadratic form***. Written explicitly, we see that

$$x^T A x = \sum_{i=1}^{n} x_i (Ax)_i = \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{n} A_{ij} x_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j \quad .$$

Note that,

$$x^T A x = (x^T A x)^T = x^T A^T x = x^T \left(\frac{1}{2} A + \frac{1}{2} A^T \right) x,$$

Operations and properties

- Positive semidefinite matrices

Positive definite matrix is always full rank

- A symmetric matrix $A \in \mathbb{S}^n$ is ***positive definite*** (PD) if for all non-zero vectors $x \in \mathbb{R}^n$, $x^T Ax > 0$. This is usually denoted $A \succ 0$ (or just $A > 0$), and often times the set of all positive definite matrices is denoted $\mathbb{S}^n_{++}$.
- A symmetric matrix $A \in \mathbb{S}^n$ is ***positive semidefinite*** (PSD) if for all vectors $x^T Ax \geq 0$. This is written $A \succeq 0$ (or just $A \geq 0$), and the set of all positive semidefinite matrices is often denoted $\mathbb{S}^n_+$.
- Likewise, a symmetric matrix $A \in \mathbb{S}^n$ is ***negative definite*** (ND), denoted $A \prec 0$ (or just $A < 0$) if for all non-zero $x \in \mathbb{R}^n$, $x^T Ax < 0$.
- Similarly, a symmetric matrix $A \in \mathbb{S}^n$ is ***negative semidefinite*** (NSD), denoted $A \preceq 0$ (or just $A \leq 0$) if for all $x \in \mathbb{R}^n$, $x^T Ax \leq 0$.
- Finally, a symmetric matrix $A \in \mathbb{S}^n$ is ***indefinite***, if it is neither positive semidefinite nor negative semidefinite — i.e., if there exists $x_1, x_2 \in \mathbb{R}^n$ such that $x_1^T Ax_1 > 0$ and $x_2^T Ax_2 < 0$.

Operations and properties

- Eigenvalues and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n\times n}$, we say that $\lambda \in \mathbb{C}$ is an ***eigenvalue*** of A and $x \in \mathbb{C}^n$ is the corresponding ***eigenvector***[3] if

$$Ax = \lambda x, \quad x \neq 0.$$

The following are properties of eigenvalues and eigenvectors (in all cases assume $A \in \mathbb{R}^{n\times n}$ has eigenvalues $\lambda_i, \ldots, \lambda_n$ and associated eigenvectors $x_1, \ldots x_n$):

- The trace of a A is equal to the sum of its eigenvalues, $\mathrm{tr} A = \sum_{i=1}^{n} \lambda_i$.
- The determinant of A is equal to the product of its eigenvalues, $|A| = \prod_{i=1}^{n} \lambda_i$.
- The rank of A is equal to the number of non-zero eigenvalues of A.
- If A is non-singular then $1/\lambda_i$ is an eigenvalue of A^{-1} with associated eigenvector x_i, i.e., $A^{-1}x_i = (1/\lambda_i)x_i$. (To prove this, take the eigenvector equation, $Ax_i = \lambda_i x_i$ and left-multiply each side by A^{-1}.)
- The eigenvalues of a diagonal matrix $D = \mathrm{diag}(d_1, \ldots d_n)$ are just the diagonal entries $d_1, \ldots d_n$.

Matrix calculus

- The Gradient and the Hessian

Suppose that $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a function that takes as input a matrix A of size $m \times n$ and returns a real value.

$$\nabla_A f(A) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

$$\nabla_x^2 f(x) \in \mathbb{R}^{n\times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

$$f(z) = z^T z$$

$$\nabla_z f(z) = 2z$$

Matrix calculus

- Derivatives of Matrices, Vectors and Scalar Forms
  - First order:

$$\frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} \qquad \frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T \qquad \frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T$$

  - Second order:

$$\frac{\partial \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{c}}{\partial \mathbf{X}} = \mathbf{X}(\mathbf{b}\mathbf{c}^T + \mathbf{c}\mathbf{b}^T) \qquad \frac{\partial \mathbf{b}^T\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{c}}{\partial \mathbf{X}} = \mathbf{D}^T\mathbf{X}\mathbf{b}\mathbf{c}^T + \mathbf{D}\mathbf{X}\mathbf{c}\mathbf{b}^T$$

$$\frac{\partial (\mathbf{B}\mathbf{x}+\mathbf{b})^T\mathbf{C}(\mathbf{D}\mathbf{x}+\mathbf{d})}{\partial \mathbf{x}} = \mathbf{B}^T\mathbf{C}(\mathbf{D}\mathbf{x}+\mathbf{d}) + \mathbf{D}^T\mathbf{C}^T(\mathbf{B}\mathbf{x}+\mathbf{b})$$

$$\frac{\partial \mathbf{x}^T\mathbf{B}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{B}+\mathbf{B}^T)\mathbf{x} \qquad \frac{\partial}{\partial \mathbf{X}}(\mathbf{X}\mathbf{b}+\mathbf{c})^T\mathbf{D}(\mathbf{X}\mathbf{b}+\mathbf{c}) = (\mathbf{D}+\mathbf{D}^T)(\mathbf{X}\mathbf{b}+\mathbf{c})\mathbf{b}^T$$

Matrix calculus

- Derivatives of Traces
  - First order

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{X}) &= \mathbf{I} \\
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{XA}) &= \mathbf{A}^T \\
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{AXB}) &= \mathbf{A}^T\mathbf{B}^T \\
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{AX}^T\mathbf{B}) &= \mathbf{BA} \\
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{X}^T\mathbf{A}) &= \mathbf{A} \\
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{AX}^T) &= \mathbf{A}
\end{aligned}$$

$$\boxed{\frac{\partial}{\partial \mathbf{X}}||\mathbf{X}||_{\mathrm{F}}^2 = \frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{XX}^H) = 2\mathbf{X}}$$

  - Second order

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{X}^2) &= 2\mathbf{X}^T \\
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{X}^2\mathbf{B}) &= (\mathbf{XB}+\mathbf{BX})^T \\
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{X}^T\mathbf{BX}) &= \mathbf{BX}+\mathbf{B}^T\mathbf{X} \\
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{XBX}^T) &= \mathbf{XB}^T+\mathbf{XB} \\
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{AXBX}) &= \mathbf{A}^T\mathbf{X}^T\mathbf{B}^T+\mathbf{B}^T\mathbf{X}^T\mathbf{A}^T \\
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{X}^T\mathbf{X}) &= 2\mathbf{X} \\
\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}(\mathbf{BXX}^T) &= (\mathbf{B}+\mathbf{B}^T)\mathbf{X}
\end{aligned}$$

Matrix calculus

- Least Squares

Suppose we are given matrices $A \in \mathbb{R}^{m \times n}$ (for simplicity we assume A is full rank) and a vector $b \in \mathbb{R}^m$ such that $b \notin \mathcal{R}(A)$.

$$\boxed{\frac{\partial \mathbf{x}^T \mathbf{B} \mathbf{x}}{\partial \mathbf{x}} = (\mathbf{B} + \mathbf{B}^T)\mathbf{x}}$$

$$\boxed{\frac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}}$$

$$\begin{aligned} \|Ax - b\|_2^2 &= (Ax - b)^T (Ax - b) \\ &= x^T A^T A x - 2b^T A x + b^T b \end{aligned}$$

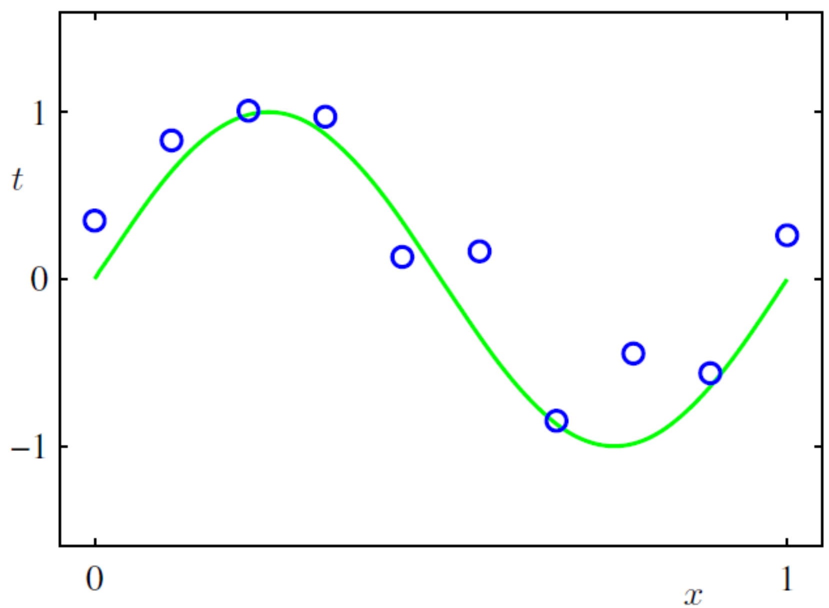
$$\begin{aligned} \nabla_x (x^T A^T A x - 2b^T A x + b^T b) &= \nabla_x x^T A^T A x - \nabla_x 2b^T A x + \nabla_x b^T b \\ &= 2A^T A x - 2A^T b \end{aligned}$$

$$x = (A^T A)^{-1} A^T b$$

Example: Polynomial curve fitting

Example: Polynomial curve fitting

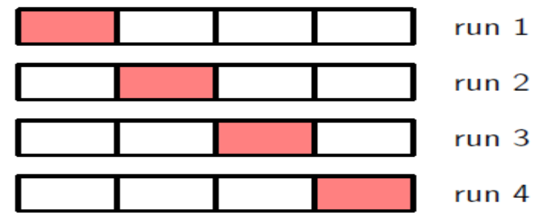
- Training data set
- Target data set
- Linear model

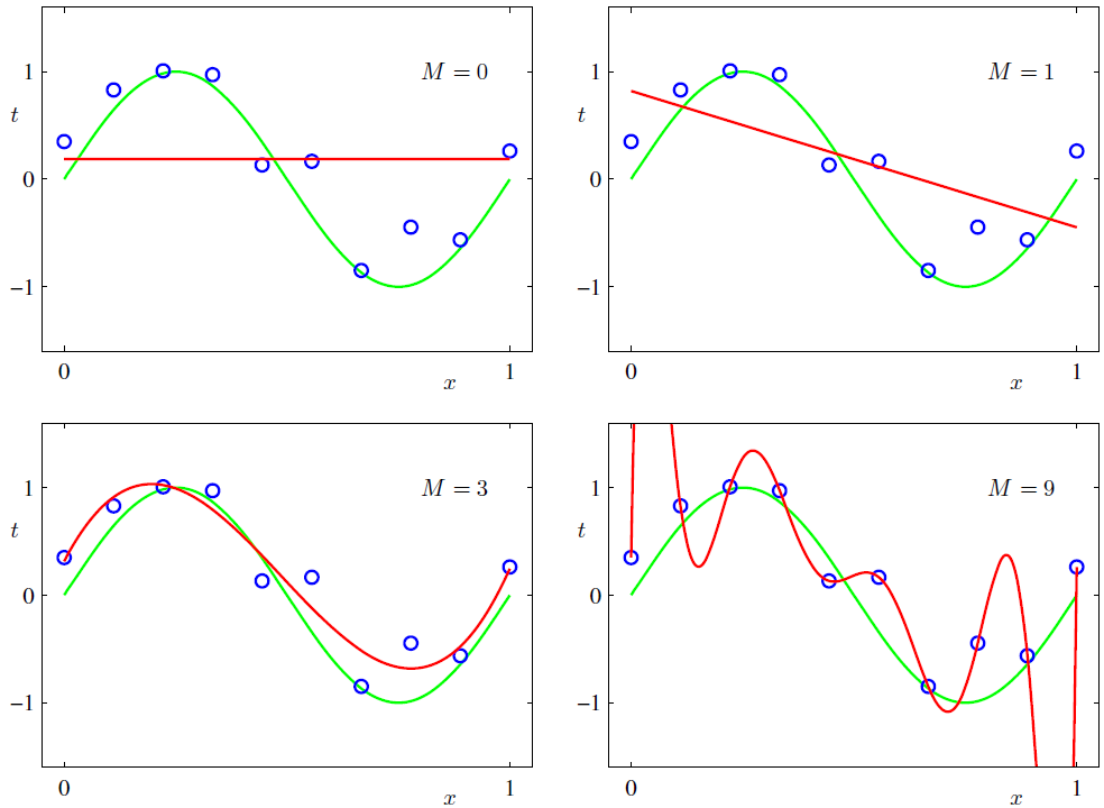


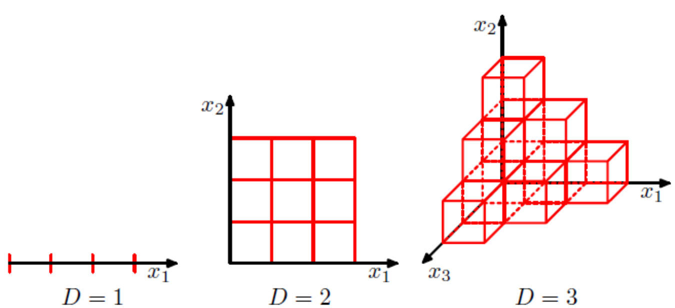
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Example: Polynomial curve fitting

- Model comparison or model selection
  - validation set, Cross-validation (CV)
- Over-fitting
  - How to control?
  - *Regularization (penalty term)*
  - *Bayesian approach (prior)*
  - *CV…*
- The curse of dimensionality



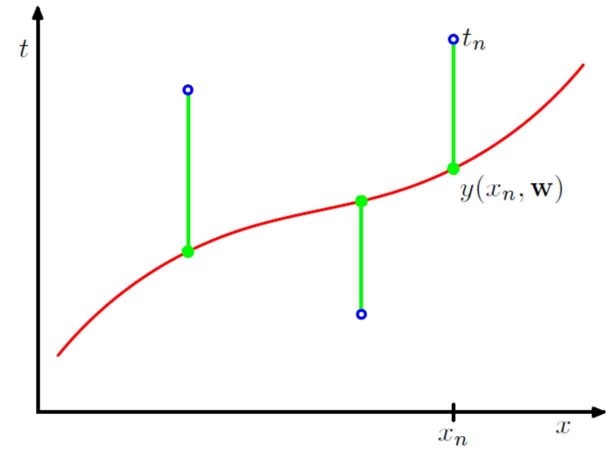




Example: Polynomial curve fitting

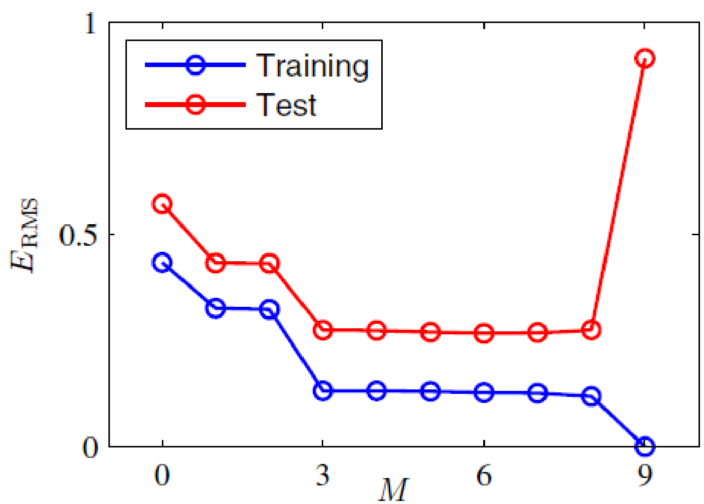
- Error function:
  - SSE (sum-of-square) error

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



  - RMS (root-mean-square) error

$$E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$$



Example: Polynomial curve fitting

- The size of the data set N vs. model parameters K

N > (5~10)* K or Bayesian approach

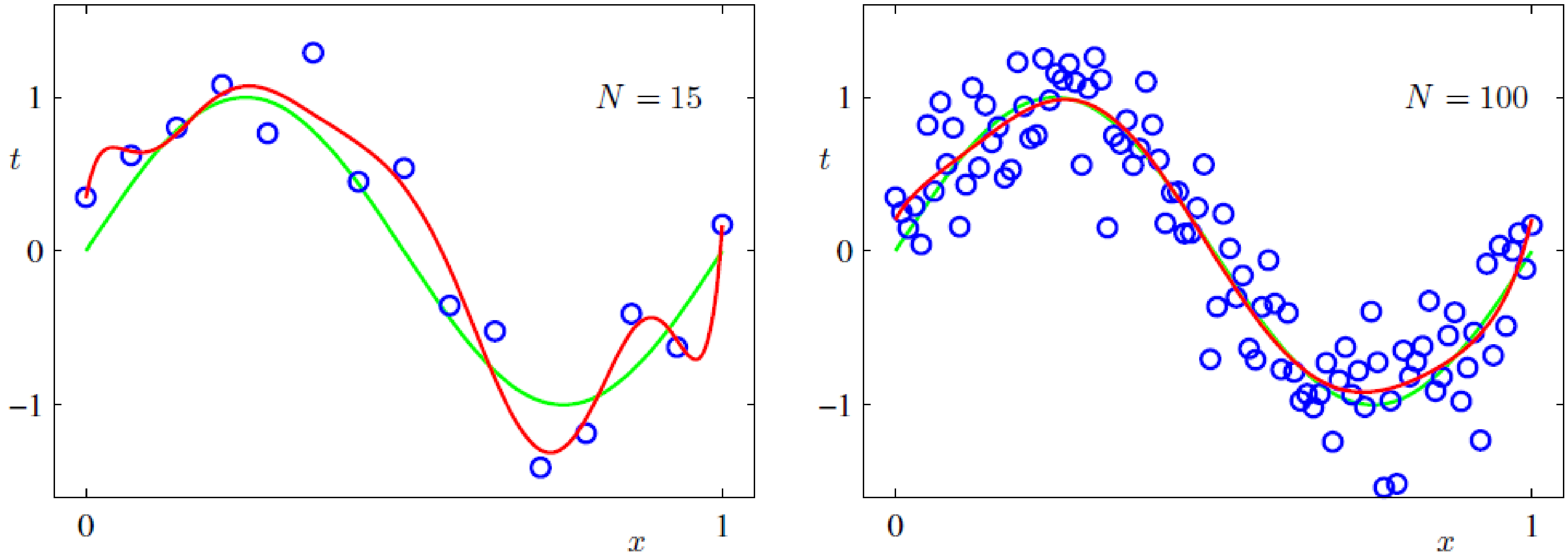


Figure 1.6 Plots of the solutions obtained by minimizing the sum-of-squares error function using the $M = 9$ polynomial for $N = 15$ data points (left plot) and $N = 100$ data points (right plot). We see that increasing the size of the data set reduces the over-fitting problem.

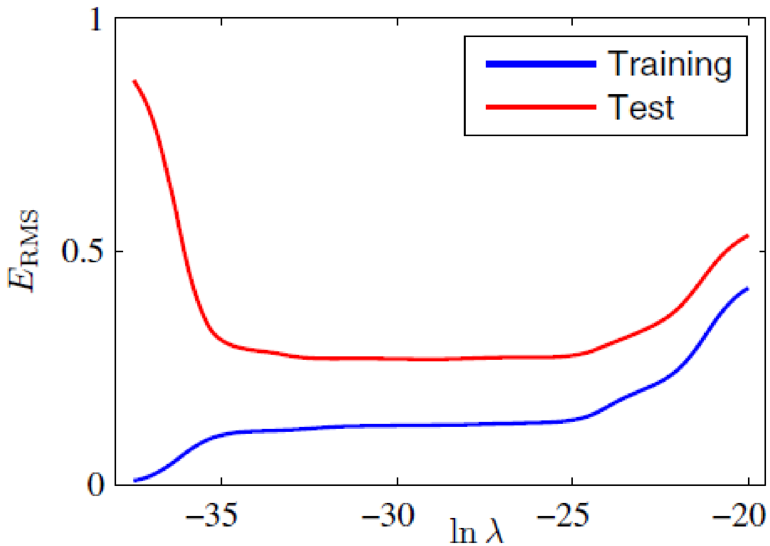
Example: Polynomial curve fitting

- Regularization
  - Penalty term
  - Closed form
  - Shrinkage methods
  - Ridge regression

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

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

	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
$w_0^\star$	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



Probability theory review and notation

Basic notation

- Random variable and rules of probability
  - Sum rule (marginal)

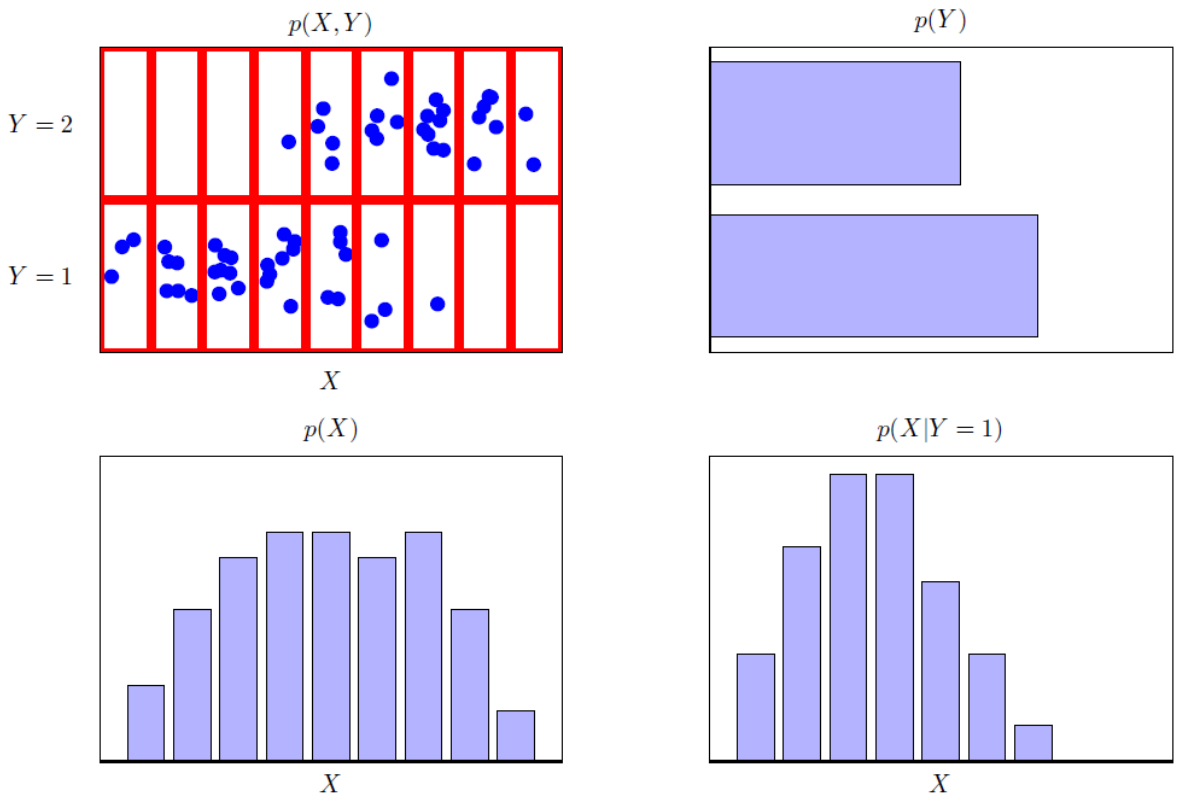
$$p(X) = \sum_{Y} p(X, Y)$$

  - Product rule (joint)

$$p(X, Y) = p(Y|X)p(X)$$

  - Bayes's theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$



Basic notation

- A piece of fruit is removed from the box (with equal probability of selecting any of the items in the box), then what is the probability of selecting an apple?

$$
\begin{aligned}
p(a) &= p(a|r)p(r) + p(a|b)p(b) + p(a|g)p(g) \\
&= \frac{3}{10} \times 0.2 + \frac{1}{2} \times 0.2 + \frac{3}{10} \times 0.6 = 0.34
\end{aligned}
$$

	Red	Blue	Green
	3	1	3
	4	1	3
	3	0	4
	p(r) = 0.2	p(b) = 0.2	p(g) = 0.6

- If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

$$
p(g|o) = \frac{p(o|g)p(g)}{p(o)}
$$

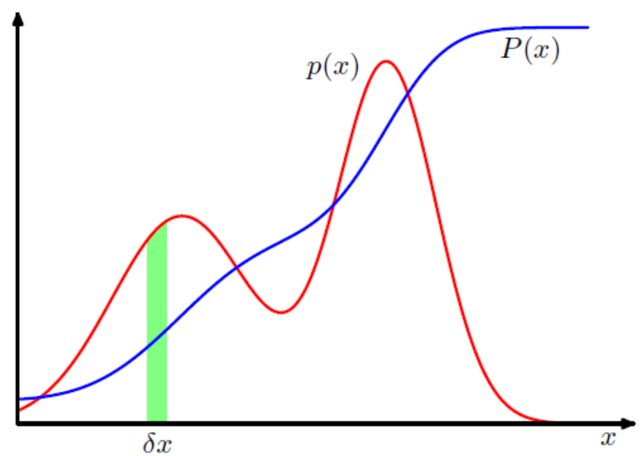
$$
\begin{aligned}
p(o) &= p(o|r)p(r) + p(o|b)p(b) + p(o|g)p(g) \\
&= \frac{4}{10} \times 0.2 + \frac{1}{2} \times 0.2 + \frac{3}{10} \times 0.6 = 0.36
\end{aligned}
$$

$$
p(g|o) = \frac{3}{10} \times \frac{0.6}{0.36} = \frac{1}{2}
$$

Probability densities

- Probability density function:

$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x$$



- *p(x)* must satisfy the two conditions:

$$\begin{aligned} p(x) &\geqslant 0 \\ \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x &= 1 \end{aligned}$$

- Cumulative distribution function

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

Expectations and covariances

- Expectation of *f(x)* under a probability distribution *p(x)*

$$\mathbb{E}[f] = \int p(x) f(x)\,\mathrm{d}x \qquad \mathbb{E}[f] = \sum_x p(x) f(x) \qquad \mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

- Multiple variables: $\mathbb{E}_x[f(x,y)] = \int p(x) f(x,y)\,\mathrm{d}x$
- Conditional expectation: $\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$
- Variance of *f(x)*: $\boxed{\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2}$
- Covariance:

$$\begin{aligned}
\mathrm{cov}[x,y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \\
\mathrm{cov}[\mathbf{x},\mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}].
\end{aligned}$$

Bayesian probabilities

- Frequentist statistics vs. Bayesian statistics
  - View probabilities in terms of the frequencies of random, repeatable events.
  - Probabilities provide a quantification of uncertainty.

Thomas Bayes
1701–1761

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})p(\mathbf{w})\,\mathrm{d}\mathbf{w}$$

- Maximum Likelihood
- Maximum posterior - MAP

The Gaussian distribution

- Gaussian distribution (normal distribution)

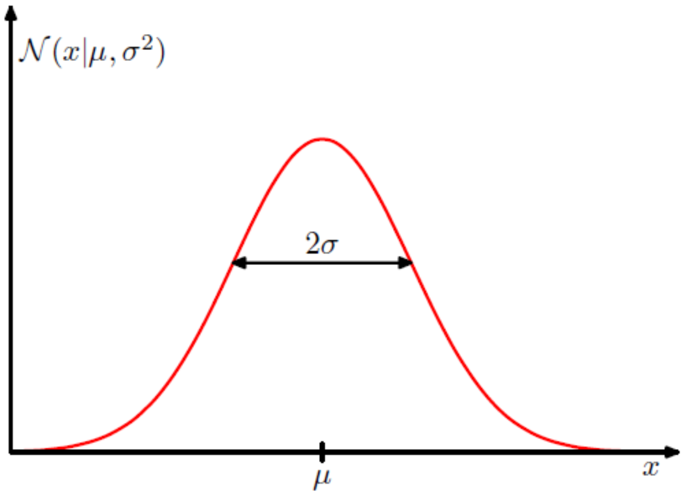
$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) \mathrm{d}x = 1$$

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

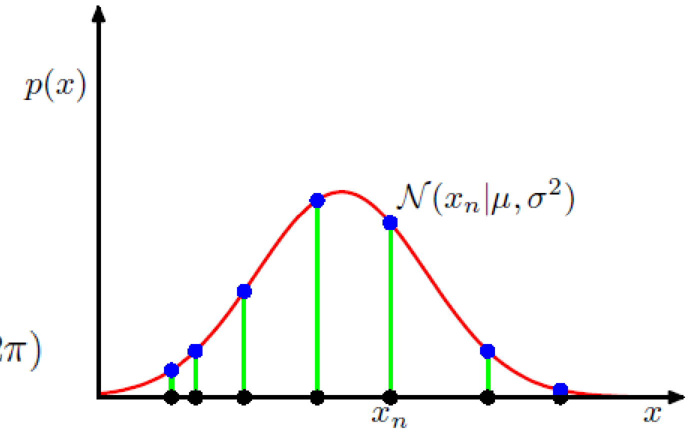


$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

The Gaussian distribution

- independent and identically distributed (i.i.d)

$$p(\mathbf{x}|\mu,\sigma^2)=\prod_{n=1}^{N}\mathcal{N}\left(x_n|\mu,\sigma^2\right)$$



- Log likelihood:

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right)=-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2-\frac{N}{2}\ln\sigma^2-\frac{N}{2}\ln(2\pi)$$

- The maximum likelihood solution:

$$\mu_{\mathrm{ML}}=\frac{1}{N}\sum_{n=1}^{N}x_n \qquad \sigma^2_{\mathrm{ML}}=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

$$\begin{aligned}\mathbb{E}[\mu_{\mathrm{ML}}] &= \mu\\ \mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \left(\frac{N-1}{N}\right)\sigma^2\end{aligned}$$

$$\widetilde{\sigma}^2=\frac{N}{N-1}\sigma^2_{\mathrm{ML}}=\frac{1}{N-1}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

Curve fitting: probabilistic perspective

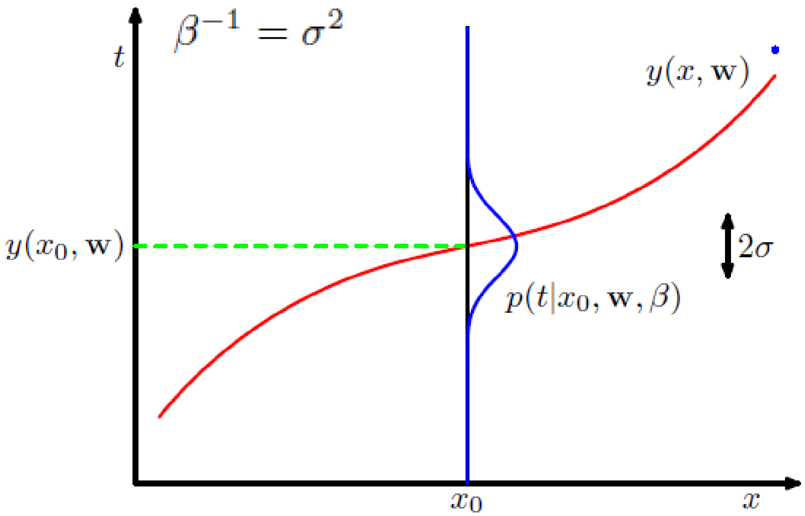
Curve fitting re-visited

- Express uncertainty over the value of the target variable using a probability distribution:

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

training data $\{\mathbf{x}, \mathbf{t}\}$

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



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

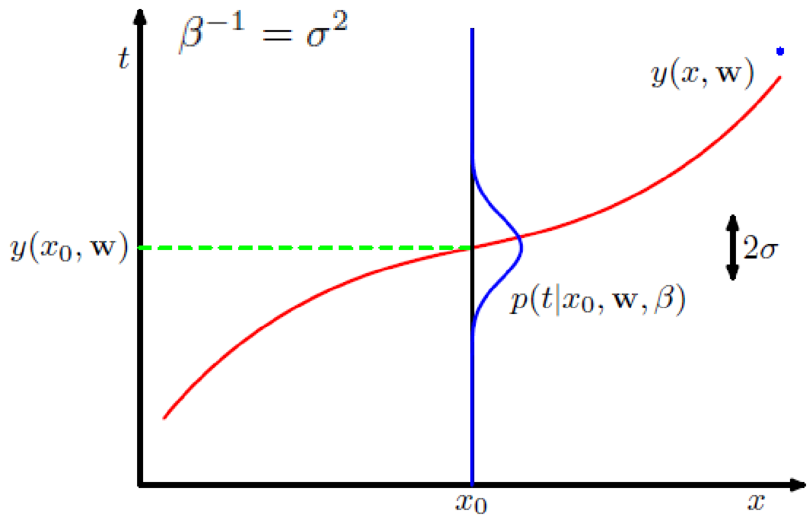
$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right)$$

Curve fitting re-visited

- More Bayesian approach:

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

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



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

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

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

maximizing the posterior distribution is equivalent to minimizing the regularized sum-of-squares error function encountered earlier in the form (1.4), with a regularization parameter given by $\lambda = \alpha/\beta$.

Bayesian curve fitting

- Full Bayesian approach

$$\boxed{p(t|x,\mathbf{x},\mathbf{t}) = \int p(t|x,\mathbf{w})p(\mathbf{w}|\mathbf{x},\mathbf{t})\,\mathrm{d}\mathbf{w}}$$

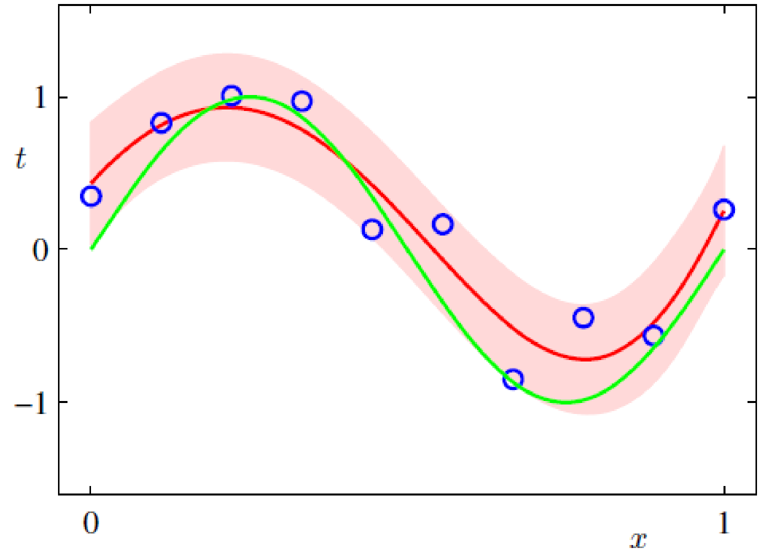
$$\boxed{p(t|x,\mathbf{x},\mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right)}$$

$$\begin{aligned} m(x) &= \beta\phi(x)^{\mathrm{T}}\mathbf{S}\sum_{n=1}^{N}\phi(x_n)t_n \\ s^2(x) &= \beta^{-1} + \phi(x)^{\mathrm{T}}\mathbf{S}\phi(x). \end{aligned}$$

$$\boxed{\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\phi(x_n)\phi(x)^{\mathrm{T}}}$$

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

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



The predictive distribution resulting from a Bayesian treatment of polynomial curve fitting using an $M = 9$ polynomial, with the fixed parameters $\alpha = 5\times10^{-3}$ and $\beta = 11.1$ (corresponding to the known noise variance), in which the red curve denotes the mean of the predictive distribution and the red region corresponds to ±1 standard deviation around the mean.

Next: Probability Distributions

- HW1:
  - PRML, Chapter-1: 1.5, 1.7~1.11
  - Submission date: In Class (16:00pm), Tuesday, March 11.