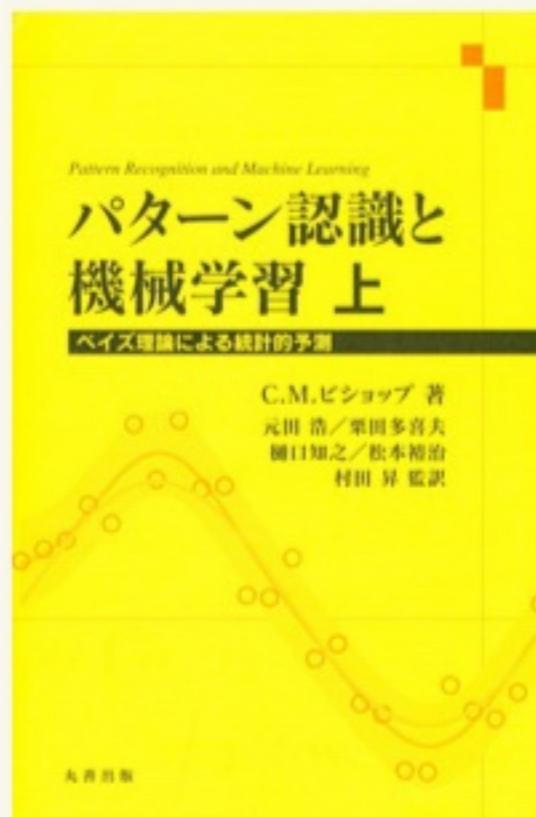


# PRML 読書会 / 第1回

@tetsuroito



## 第1章 序論

- 1.1 例：多項式曲線フィッティング
- 1.2 確率論
  - 1.2.1 確率密度
  - 1.2.2 期待値と分散
  - 1.2.3 ベイズ確率
  - 1.2.4 ガウス分布
  - 1.2.5 曲線フィッティング再訪
  - 1.2.6 ベイズ曲線フィッティング
- 1.3 モデル選択
- 1.4 次元の呪い



# 序論

データに潜むパターンを見つけ出すのは根源的なもの

16世紀：天体データからの量子力学の発見

20世紀：量子スペクトルの規則性の発見

本書の目的

計算機アルゴリズムを通じて、データの中の規則性を自動的に見つけ、  
その規則性を用いてデータを異なるカテゴリに分類する

例：手書き数字の認識



28×28ピクセルの画像

784次元の実数値ベクトルx



→xを入力として受け取り、  
0~9のどの数字かを分類する



# 序論

用語の説明（初回なので、念のため）

学習（訓練）段階：関数 $y(x)$ を訓練データで求める

テスト集合：学習されたモデルで分類する新たなデータ群

汎化：訓練で使ったのは異なる事例を分類する能力

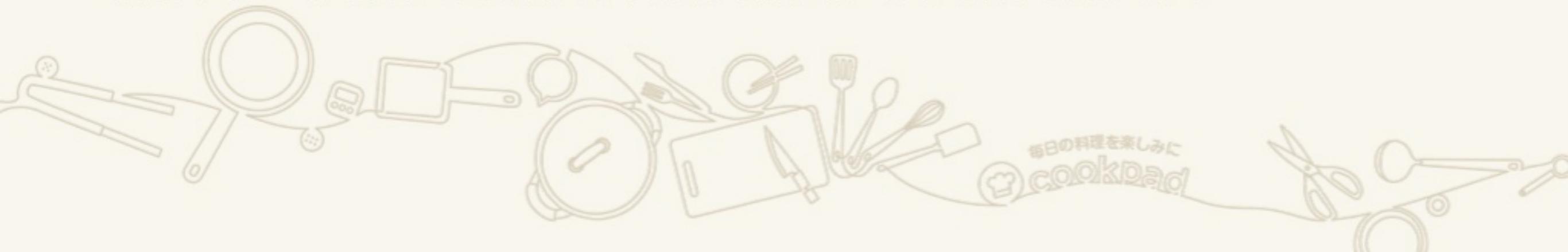
前処理：問題を解きやすくするために入力変数を変換する

特徴抽出：数字画像を平行移動、拡大縮小して固定した箱に収まるようにする  
→効率的に処理し、かつ重要な情報をそぎ落とさないように配慮する

教師あり学習：訓練データが入力および目標ベクトルに対応する問題

教師なし学習：訓練データが入力ベクトル $x$ のみ

強化学習：与えられた条件化で報酬を最大にする行動を見つける



# 1.1 多項式曲線フィッティング

N個の観測値から新たな入力値xに対して、目標変数tを予測したい

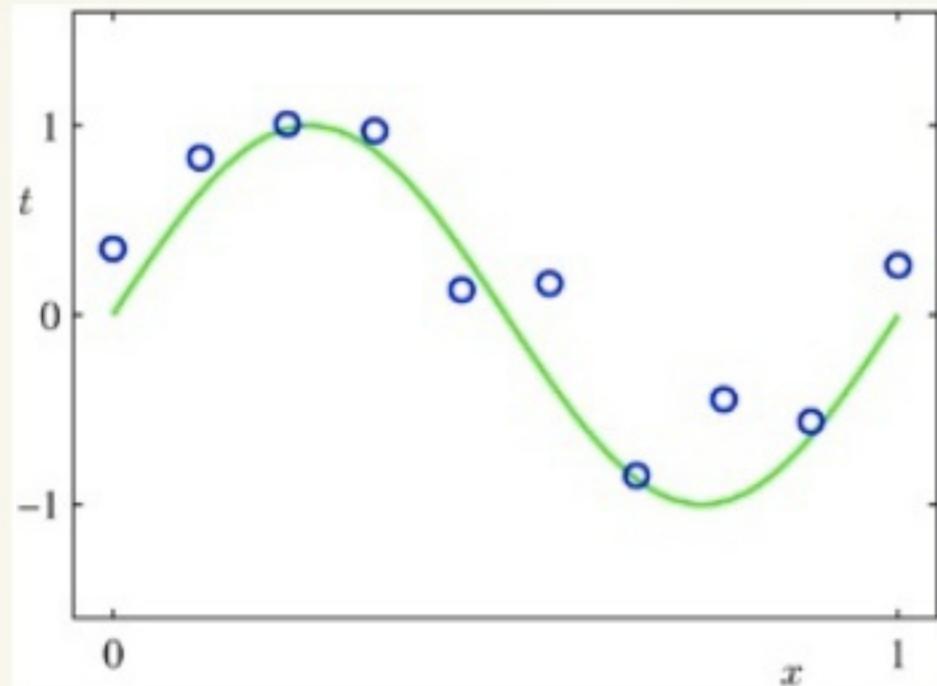


図1.2の例、N=10個のデータ集合  
生成には $\sin(2\pi x)$ +ガウスノイズ  
→曲線フィッティングで予測する  
左下のような多項式

多項式の次数

---

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

線形モデルに関しては、3章と4章で！



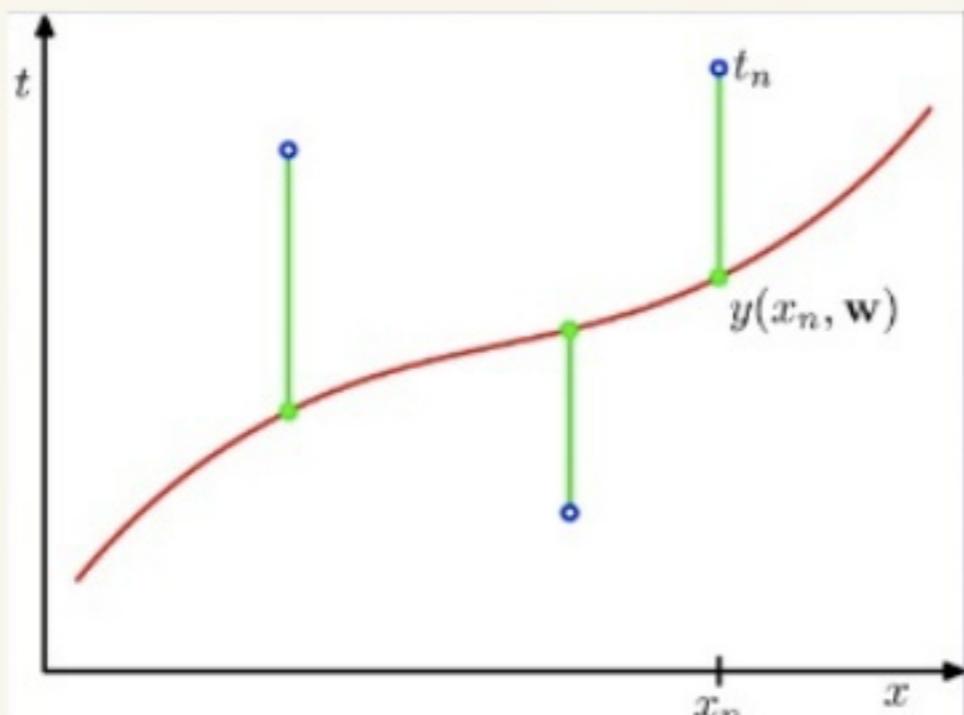
# 1.1 多項式曲線フィッティング

フィッティングをするためには…

wを任意に固定したときの関数y(x,w)の値と訓練集合のデータ点のズレを測る誤差関数を最小化する



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

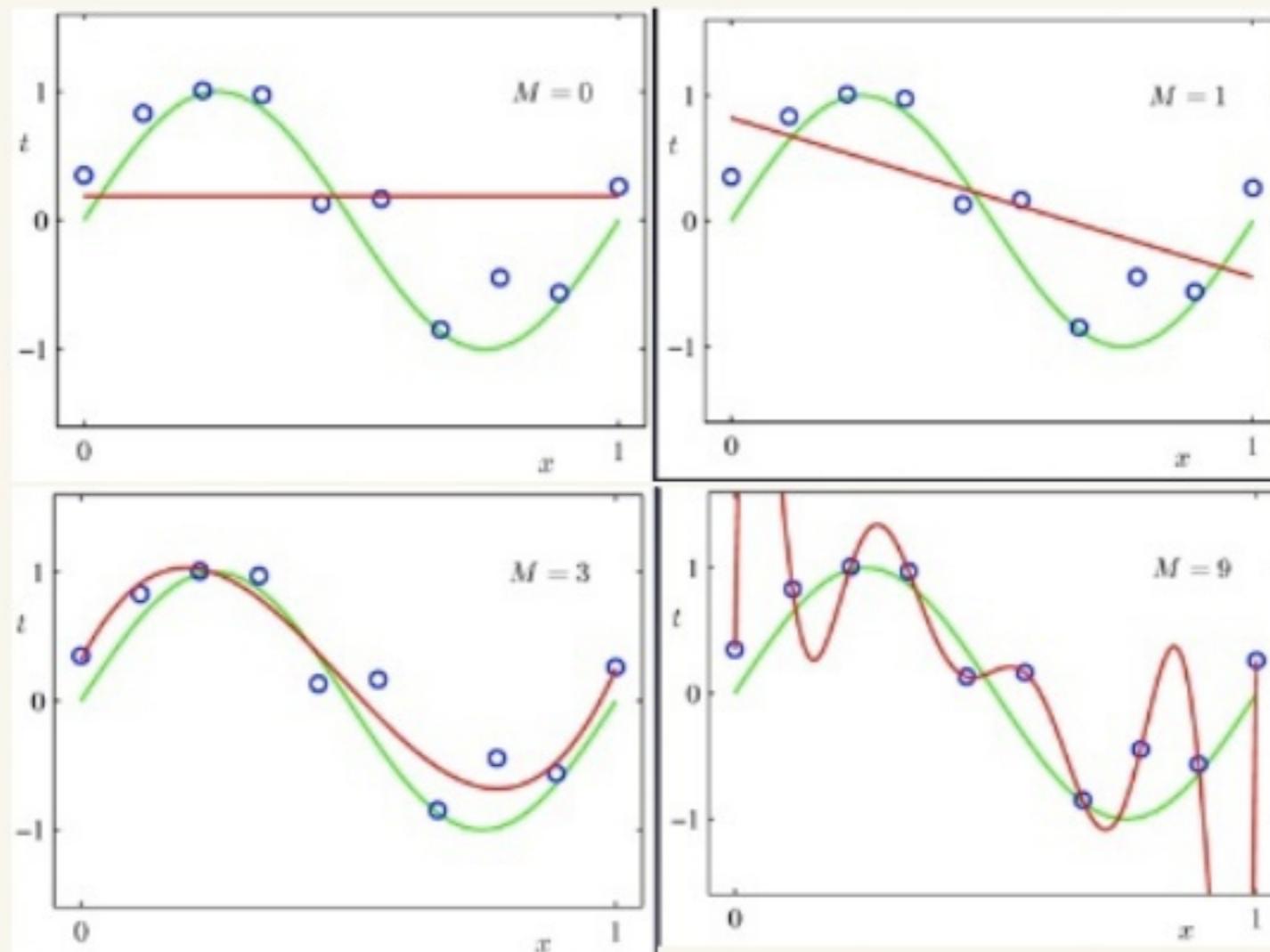


←二乗誤差関数の幾何的な解釈

なぜ、1/2をかけているか？→誤差関数が2次関数で微分したら都合がいいから  
(線形なので、解が一意)

# 1.1 多項式曲線フィッティング

モデル選択：多項式の次数を決定する



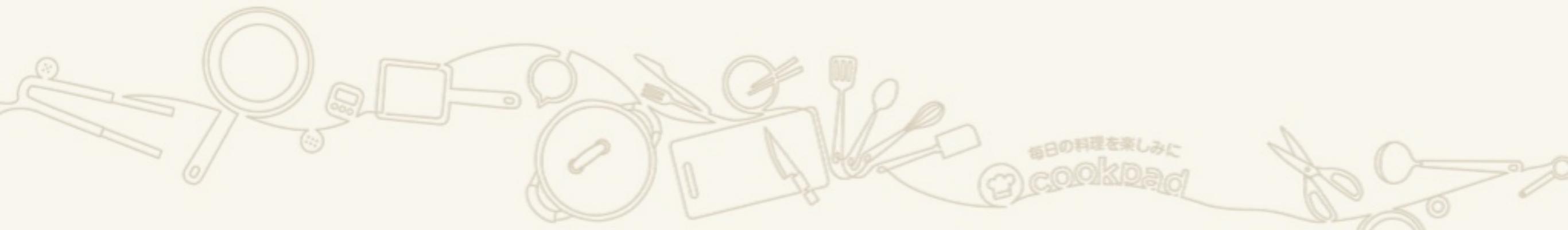
様々な次数でのフィッティング

0次や1次はあてはまり悪い

次数が高い9次（右下）

あてはまりはいい（誤差=0）が、  
過学習

3次の場合が、最もあてはまりが  
よさそう



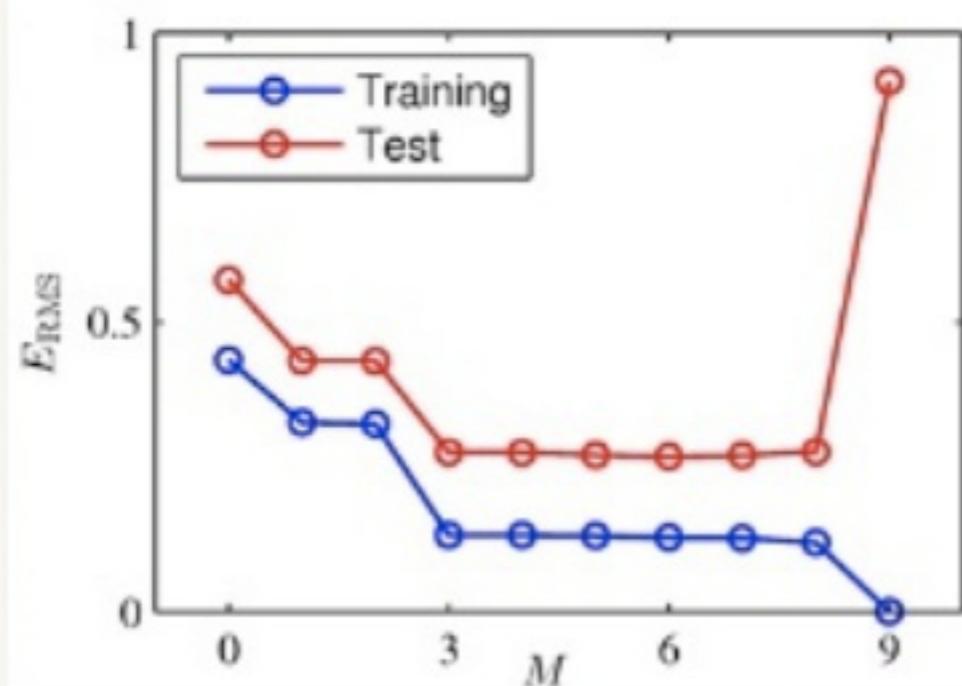
# 1.1 多項式曲線フィッティング

次数Mによる汎化性能の評価

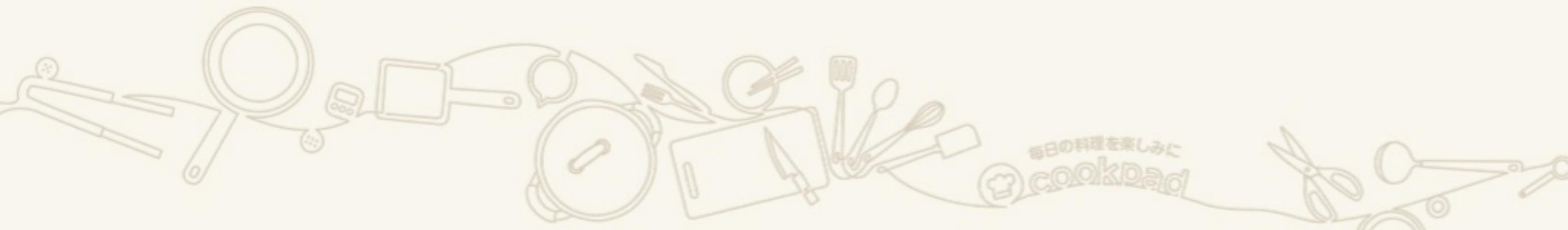
平均二乗平方根誤差を用いる

$$E_{\text{RMS}} = \sqrt{\frac{2E(\mathbf{w}^*)}{N}} = \sqrt{\sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 / N}$$

定義域： $0 \leq M \leq 9$ における訓練集合とテストデータの誤差プロット



$M=9$ は訓練データの誤差は0  
しかし、テストデータの誤差は大きい  
 $3 \leq M \leq 8$ の誤差はほぼ同じ  
よって、 $M=3$ が一番よいとなる



# 1.1 多項式曲線フィッティング

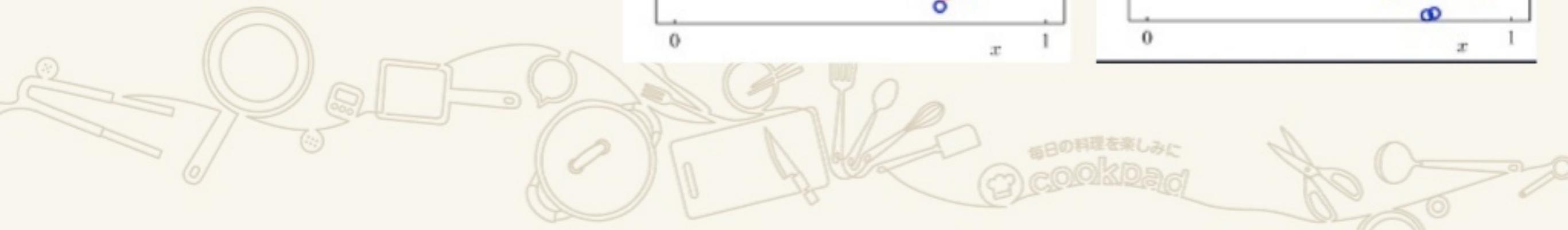
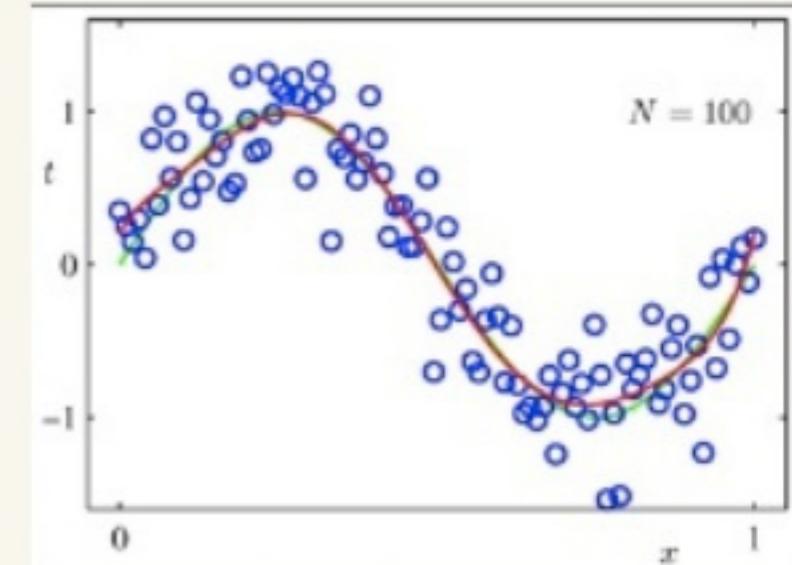
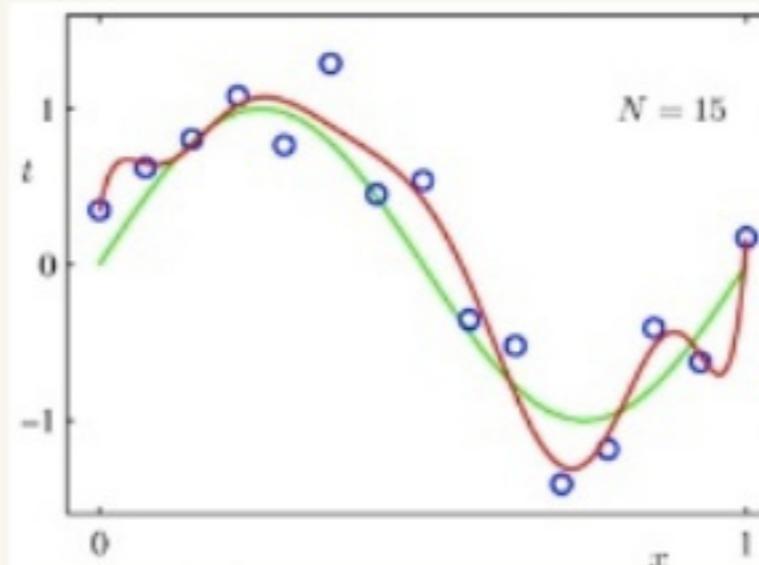
なぜ  $M=9$  の時にうまくいかないのか（同等の精度出ないの？）

様々な次数の  $w$  の係数の結果を見るとわかる

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

$M$  の増加に伴って  $w$  の係数が増大  
ランダムノイズに引きづられてしまう

データ集合のサイズを増やすと  
過学習問題はちょっと緩和



# 1.1 多項式曲線フィッティング

最尤推定において、過学習は一般的な性質

過学習を避ける方法

1、ベイズ的アプローチを用いる

(有効パラメータ数も自動的にデータ集合のサイズに適合)

2、誤差関数に罰則項を用いる (正則化)

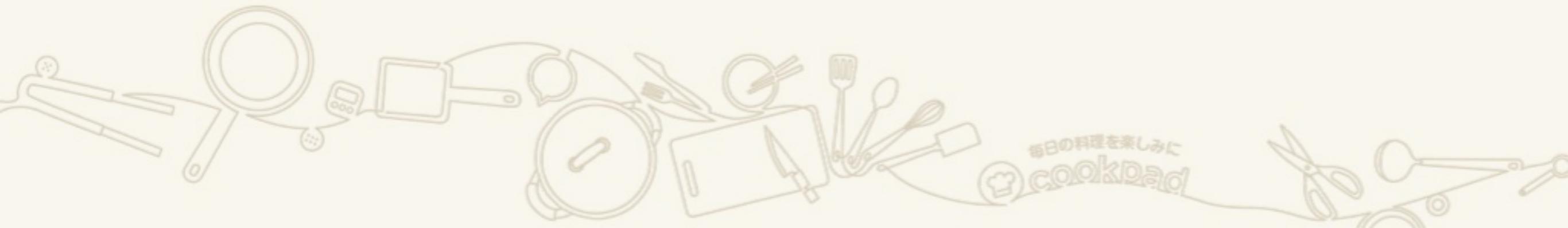
$$\tilde{E}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

係数 $\lambda$ が調整パラメータ

正則化項が2次の場合：リッジ回帰 (L2)

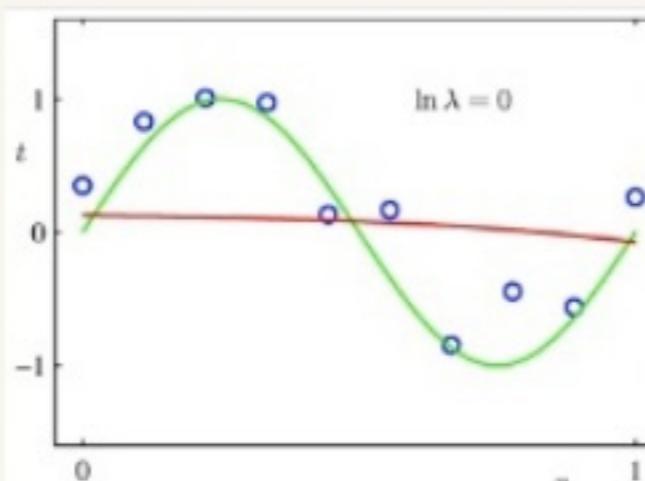
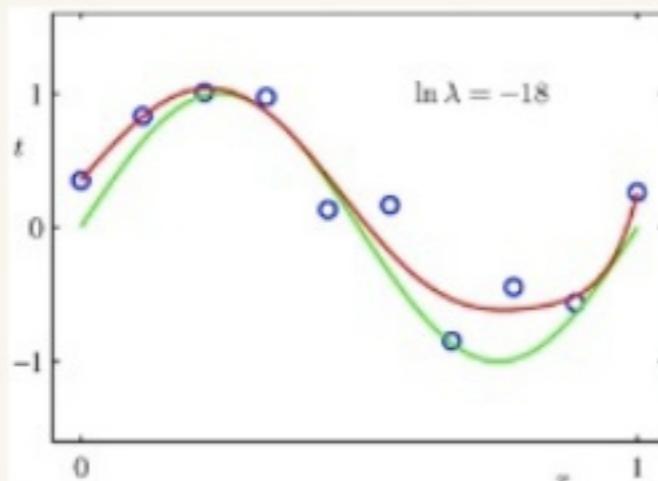
1次の場合：ラッソ回帰 (L1)

ニューラルネットワークでは荷重減衰とよばれるらしい

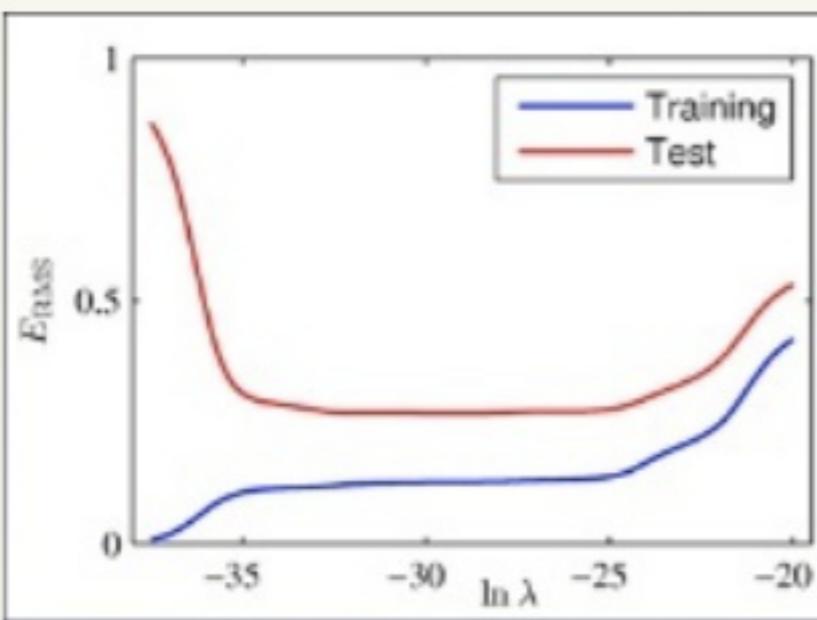


# 1.1 多項式曲線フィッティング

$\lambda$ がモデルの実質的な複雑さを制御し、過学習の度合いを決定する様



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

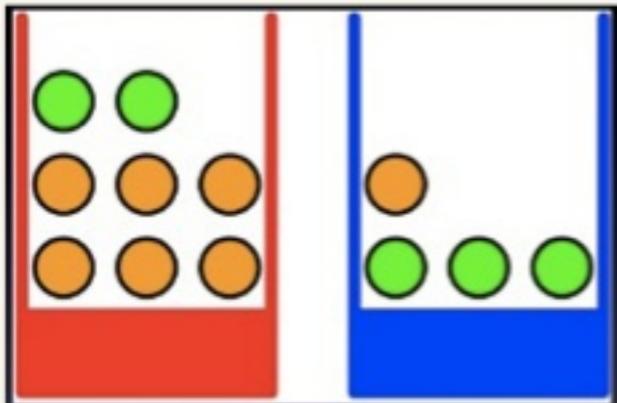


RMS誤差の値をプロット  
→訓練集合と確認集合に分ける方法  
(ホールドアウト)  
ただし、データ数との兼ね合いも考慮  
クロスバリデーションとかにつづく！



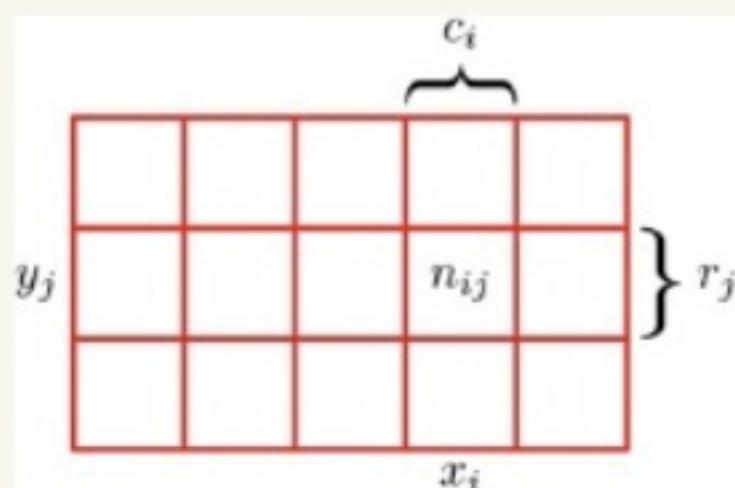
# 1.2 確率論

パターン認識における鍵となる概念：不確実性



$P(\text{赤箱}) = 4/10 \quad P(\text{青箱}) = 6/10$   
(緑：リンゴ、オレンジ：オレンジ)

一般化した設定



$M=5, L=3$  総数Nのうち、

$X=x_i, Y=y_j$  であるものを  $n_{ij}$

同時確率

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

周辺確率

$$p(X = x_i) = \frac{c_i}{N}$$

条件付き確率

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

確率の基本法則

$$p(X) = \sum_V p(X, Y)$$

加法定理

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

乗法定理

# 1.2 確率論

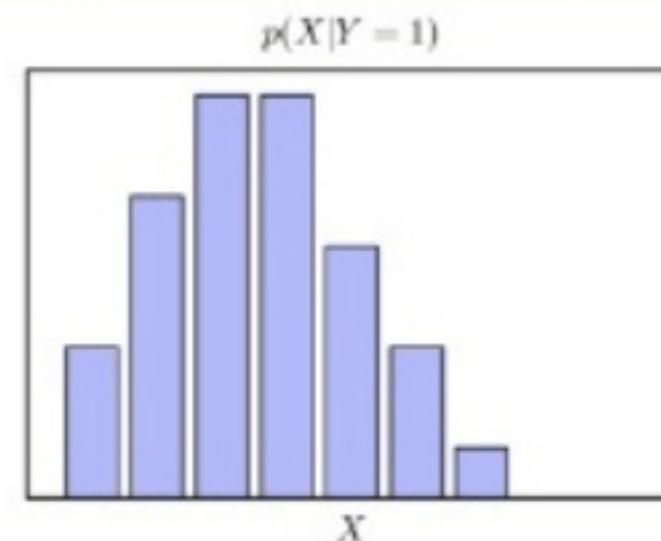
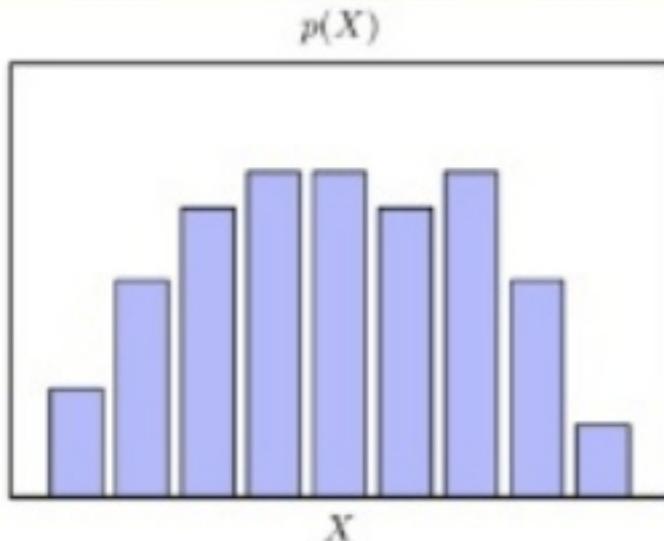
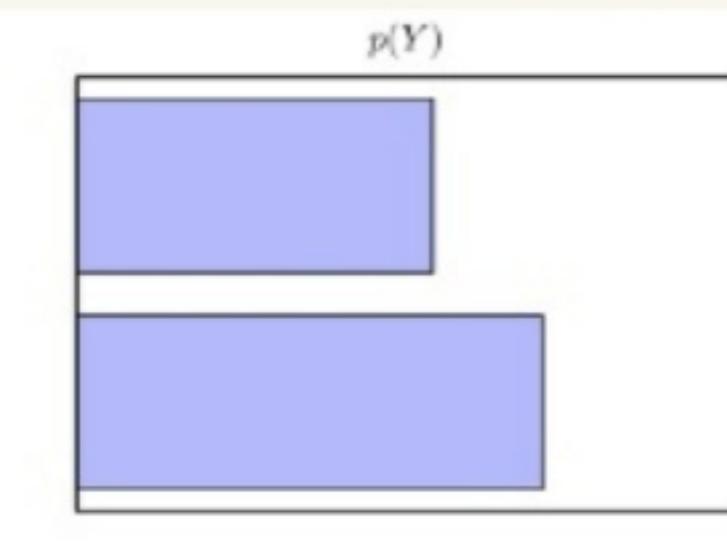
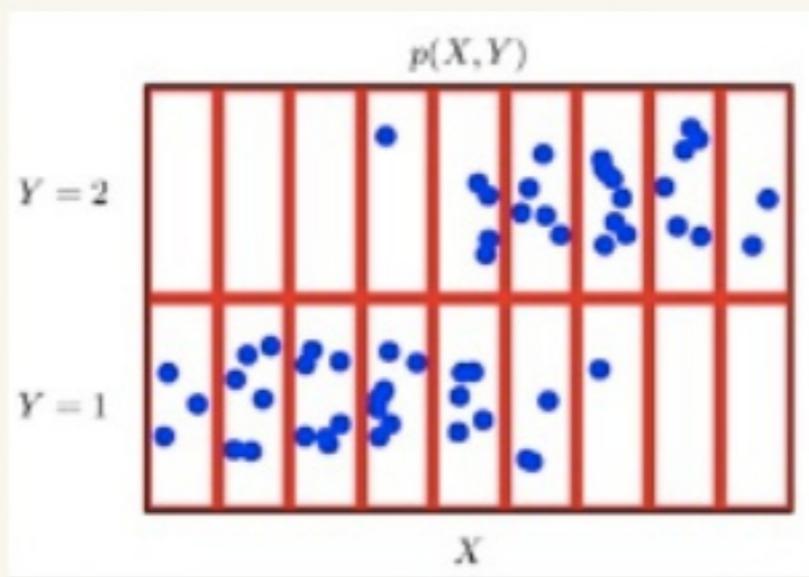
ベイズの定理

$$p(Y|X) = \frac{\text{尤度} \cdot \text{事前確率}}{\text{規格化定数}}$$
$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

事後確率

規格化定数

N=60個のプロットとヒストグラム、取りうる値など…



# 1.2 確率論

果物の例

箱を選ぶ確率

$$P(B=r) = 4/10 \quad P(B=b) = 6/10$$

果物の確率

$$P(F=a | B=r) = 1/4 \quad P(F=o | B=r) = 3/4$$

$$P(F=a | B=b) = 3/4 \quad P(F=o | B=b) = 1/4$$

(りんごを選ぶ確率)

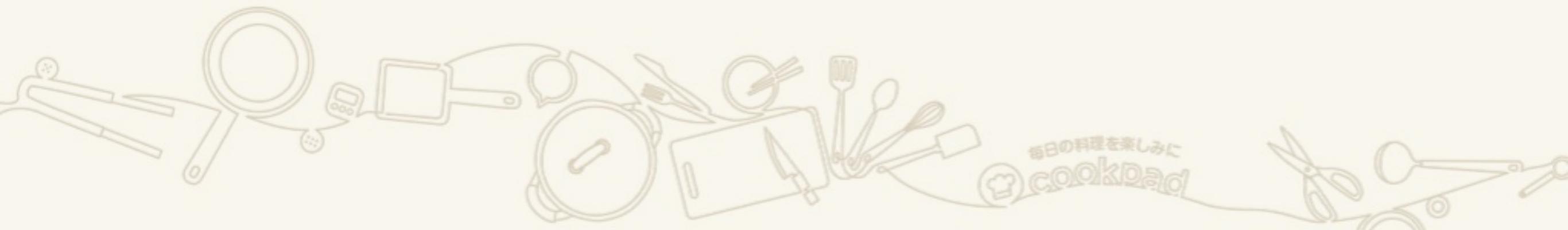
加法・乗法利用

$$P(F=a) = P(F=a|B=r)P(B=r) + P(F=a|B=b)P(B=b) = 11/20$$

(選ばれた箱の確率)

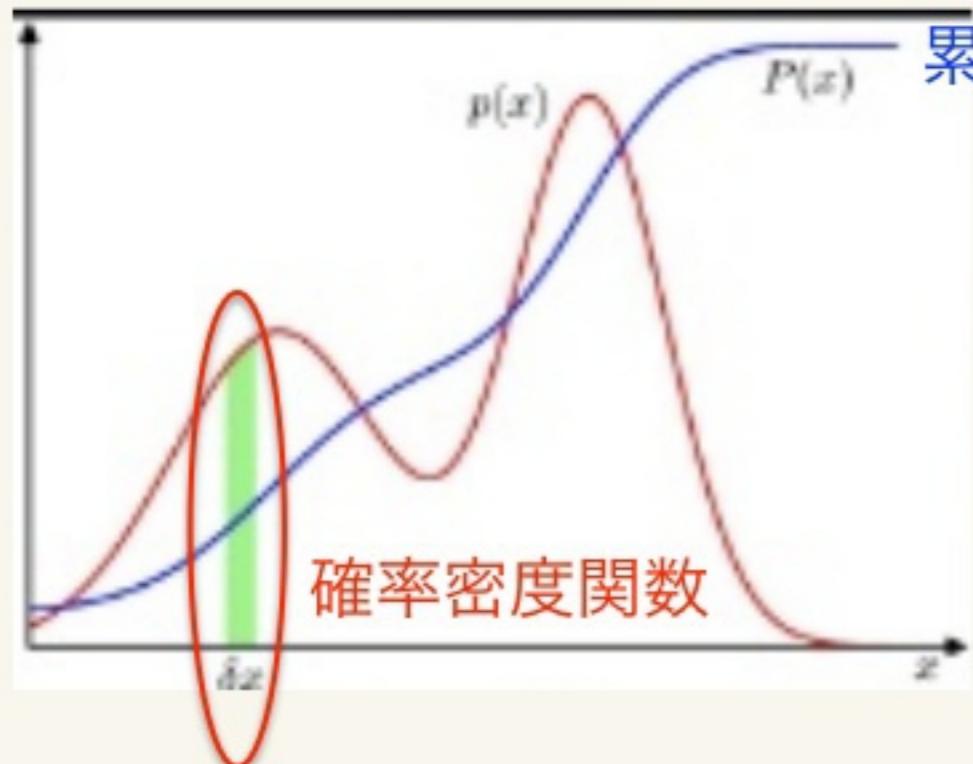
ベイズの定理利用

$$P(B=r|F=o) = \frac{P(F=o|B=r)P(B=r)}{P(F=o)} = 2/3$$



# 1.2.1 確率密度

連続変数について確率を考えた場合のもの



累積密度関数

実数値をとる変数  $x$  が区間  $(x, \delta x)$  に入る確率が  
 $\delta x \rightarrow 0$  のとき、 $p(x)\delta x$  で与えられる時の  $p(x)$

$x$  が区間  $(a, b)$  にある確率は積分で OK

$$p(x \in (a, b)) = \int_a^b p(x) dx$$

\*確率密度が満たす条件

$$p(x) \geq 0$$

確率は非負

$$\int_{-\infty}^{\infty} p(x) dx = 1$$

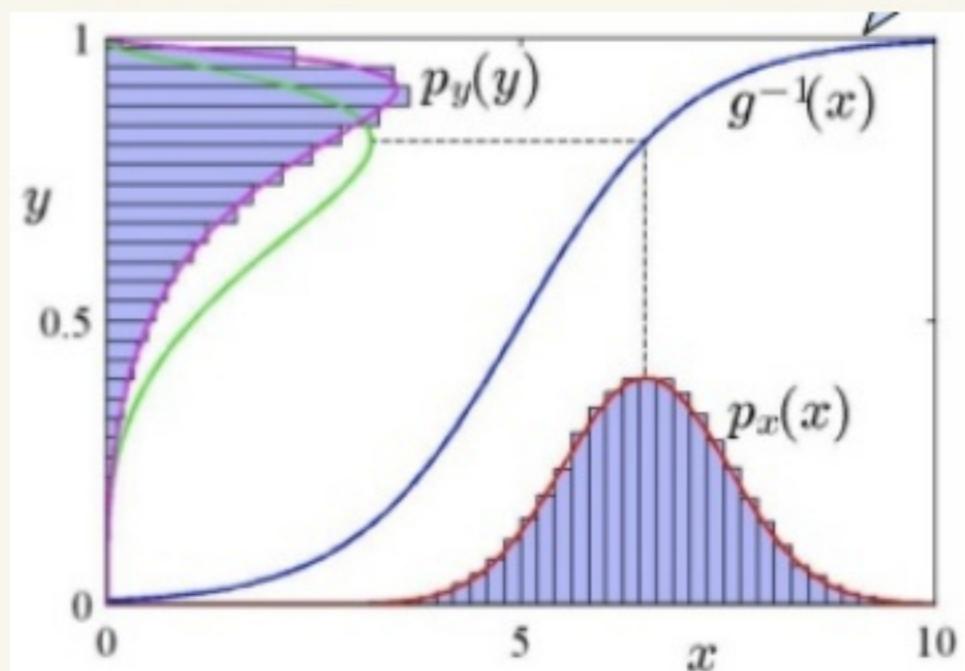
$x$  は実数上のどこかの値をとらなければならない



# 1.2.1 確率密度

分布に非線形な変換を施す

変数変換 :  $x = g(y)$  (下図青線)



ヤコビ行列により変換

$$\begin{aligned} p_y(y) &= p_x(x) \left| \frac{dx}{dy} \right| \\ &= p_x(g(y)) |g'(y)| \end{aligned}$$

確率密度の最大値は変数の選び方に依存  
要は置換積分のようですが、文系にはきつい！

累積分布関数

$x$ が区間 $(-\infty, z)$ に入る確率

$P'(x) = p(x)$ を満たす

(微分で表現)

$$P(z) = \int_{-\infty}^z p(x) dx$$



# 1.2.1 確率密度

いくつかの連続変数があるとき、まとめてベクトル $\mathbf{x}$ で表すと  
同時分布 $p(\mathbf{x})=p(x_1, \dots, x_D)$ で定義できて、 $\mathbf{x}$ を含む確率は $p(\mathbf{x})\delta\mathbf{x}$   
多変数確率密度は下記を満たす必要

$$\left| \begin{array}{l} p(\mathbf{x}) \geq 0 \\ \int p(\mathbf{x})d\mathbf{x} = 1 \end{array} \right.$$

$\mathbf{x}$ が離散の時、 $p(\mathbf{x})$ は確率質量関数とも

加法・乗法定理、ベイズの定理も適用可能

$$p(x) = \int p(x, y)dy$$

厳密に示すには測度論が必要となる

$$\underline{p(x, y) = p(y|x)p(x)}$$



# 1.2.2 期待値と分散

確率を含む最も重要な操作は関数の重み付きの平均を求めること

期待値：確率分布  $p(x)$  の下での平均値  $f(x)$

離散分布の平均

$$\mathbb{E}[f] = \sum_x p(x)f(x)$$

連続分布の平均

$$\mathbb{E}[f] = \int p(x)f(x)dx$$

いずれの場合も有限個の  $N$  点での有限和で近似可能  
(II章のサンプリング法で使うよ)

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^N f(x_n)$$

条件付き期待値

$$\mathbb{E}[f|y] = \sum_x p(x|y)f(x)$$

$$\text{var}[f] = \mathbb{E} [(f(x) - \mathbb{E}[f(x)])^2]$$

$f(x)$  の分散

$$\text{var}[f] = \mathbb{E}[(f(x)^2)] - \mathbb{E}[f(x)]^2$$

$$\text{cov}[x, y] = \mathbb{E}_{x,y} [\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}]$$

$$= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

共分散

ベクトルの共分散は教科書見て

# 1.2.3ベイズ確率

ランダムな繰り返し試行の頻度：古典的確率、頻度主義

ベイズ的な見方の導入：不確実性を定量的に表現し、新たな証拠に照らして修正、その結果として最適な行動や決定をくだす。

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

Dを観測した事後にwに関する不確実性を  
事後分布  $p(\mathbf{w}|\mathcal{D})$  の形で評価することが可能になる

→ 事後確率  $\propto$  尤度  $\times$  事前確率 【尤度の定義より】

ベイジアンと頻度主義の2つのパラダイム

尤度関数  $p(\mathcal{D}|w)$  の扱い

頻度主義：wは固定化されたパラメータで推定量  
推定の誤差範囲は D の分布を考慮  
ex) ブートストラップ法

ベイジアン：ただ1つのデータ集合 D があって、  
パラメータに関する不確実性は w の確率分布



# 1.2.3ベイズ確率

ベイズ確率への批判

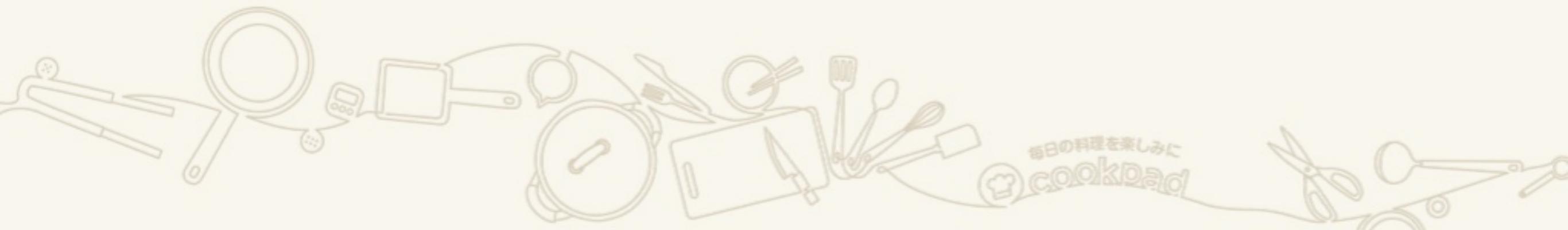
事前分布の選定が数学的な便宜や主観的になることについて

→ 無情報事前分布を用いる  
(けっこう困難な場合がある)

悪いものを選ぶと相応になるが、頻度主義の評価方法（交差検定）  
を使うと回避できる

近年はベイズが重要

全パラメータの周辺化が必要；MCMCや計算機の改良で現実的になった  
さらに、変分ベイズ法やEP法によっても有用となった



# 1.2.4 ガウス分布

みんな大好き正規分布！

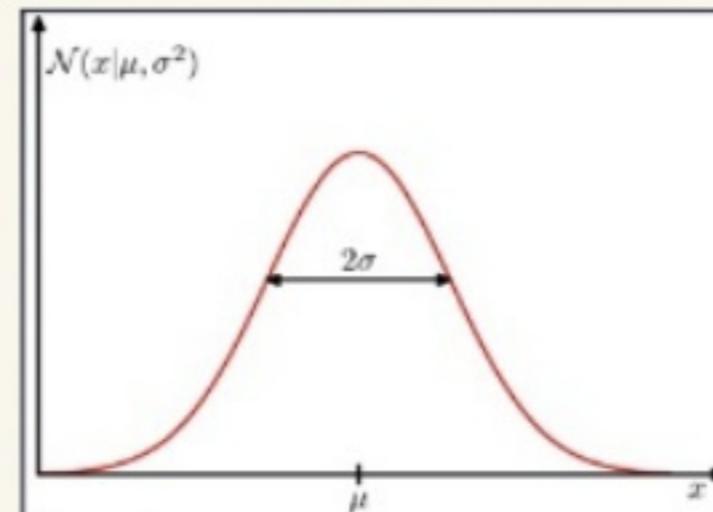
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

平均=μ 分散=σ<sup>2</sup> 標準偏差=σ 分散の逆数は精度パラメータ

満たす要件と分布

$$\mathcal{N}(x|\mu, \sigma^2) > 0$$

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



平均と2次のモーメント、分散

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

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

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

分布の最大値を与える最頻値（モード）は平均と一致

# 1.2.4 ガウス分布

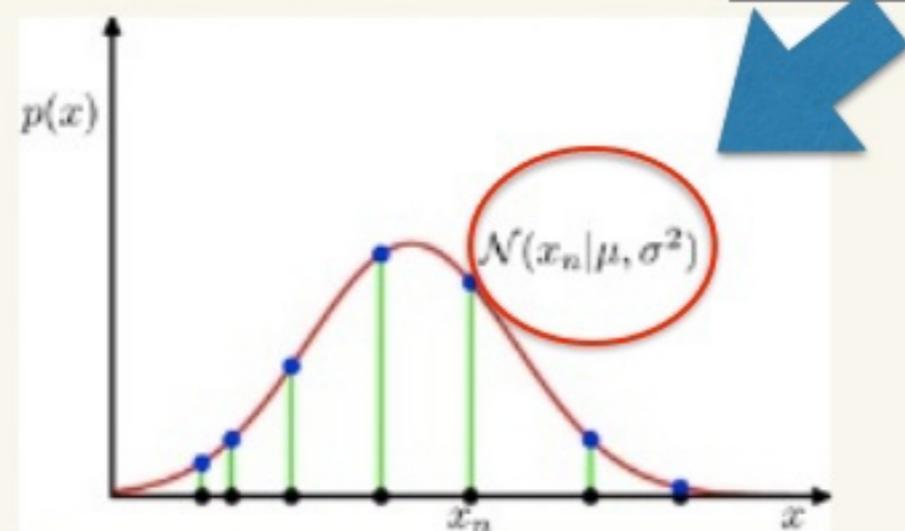
D次元ベクトルの連続変数 $\mathbf{x}$ に対するガウス分布

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

データ点が同じ分布から独立に生成される → i.i.d (独立同分布)

データ集合の確率

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



観測されたデータからパラメータ求める  
(尤度関数最大化)

便利なので、対数尤度を最大化

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

尤度関数を  $\boldsymbol{\mu}, \sigma^2$  で微分して解く

サンプル平均

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n$$

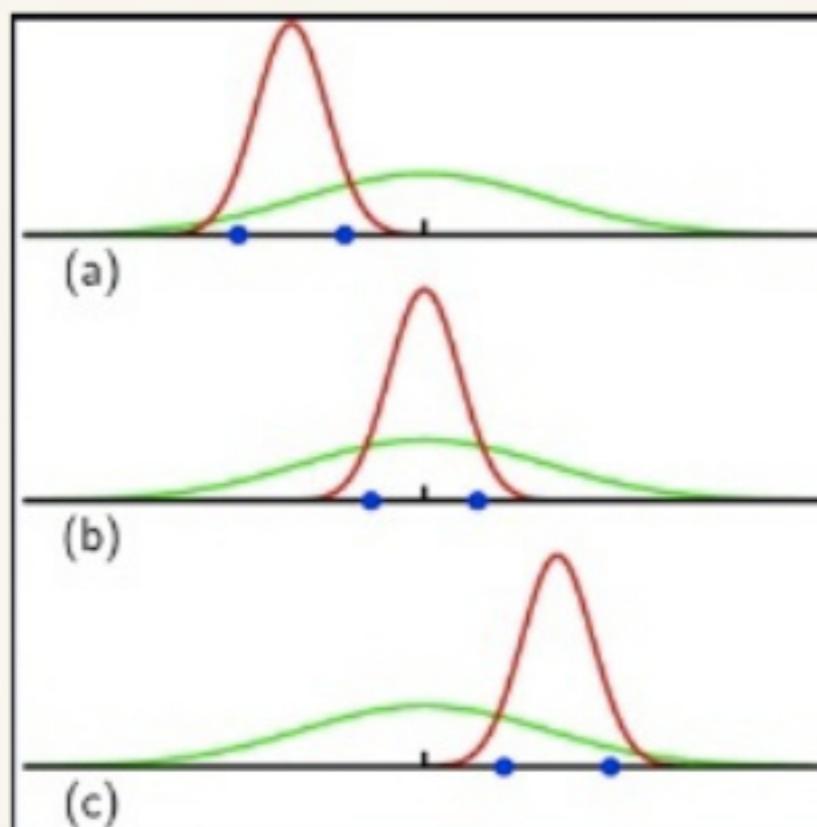
サンプル分散

$$\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})^2$$

最尤解を得る！

# 1.2.4 ガウス分布

バイアスが起きてしまう現象



緑の曲線：真のガウス分布

赤の曲線：データ集合から求めたガウス分布

青の点：データ点

平均は合っているものの、分散が過小評価されてしまうため、発生

観測データNを $\rightarrow\infty$ にすると、バイアス問題は問題ではなくなる

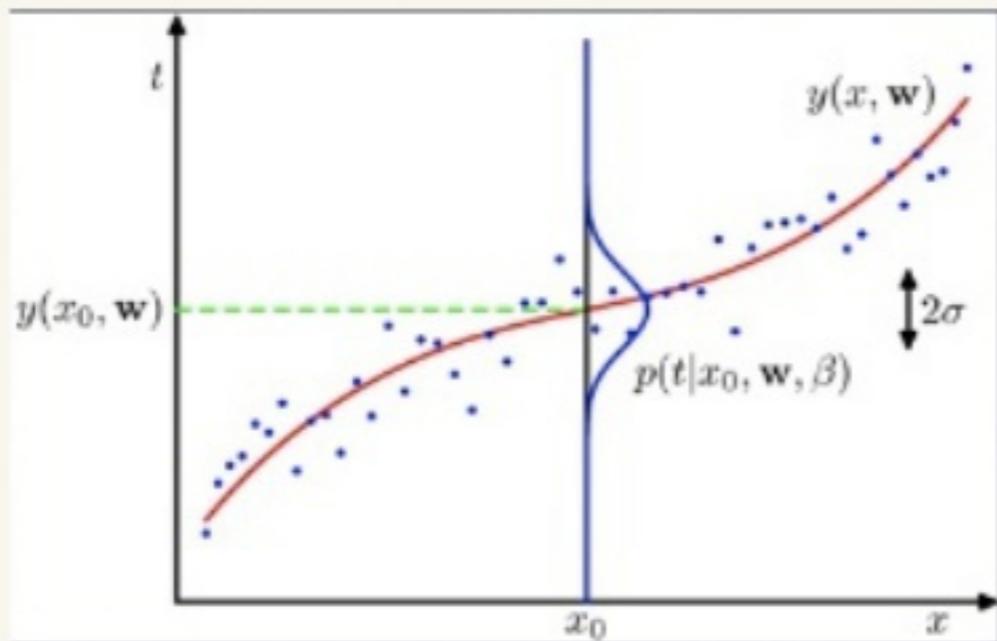


# 1.2.5 曲線フィッティング再訪

訓練データ  $\mathbf{x} = (x_1, \dots, x_N)^T$  と目標値  $\mathbf{t} = (t_1, \dots, t_N)^T$

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

精度パラメータ  $\beta$  を導入



訓練データ  $(\mathbf{x}, \mathbf{t})$  を用いて  
未知のパラメータ  $\mathbf{w}, \beta$  を求めるのに  
最尤推定を使う

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

対数尤度を最大化する ( $\mathbf{w}$ について)

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



$\mathbf{w}$ 以外を無視できるので、二乗和誤差の最小化と等価

$\beta$ についての最尤推定

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

# 1.2.5曲線フィッティング再訪

$w, \beta$ が求められたので、新たな値の予測が可能

予測分布という形での $t$ の確率分布

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

多項式 $w$ に関する事前分布を導入（ベイズ的アプローチ）

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

$\alpha$  : 精度パラメータ（超パラメータ）  
 $M+1$  :  $M$ 次多項式に対する $w$ の要素数

$w$ の事後分布は事前分布と尤度関数との積に比例する

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

→ 尤もらしい $w$ を求める → 事後分布を最大化する $w$ を求める(**MAP推定**)

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

左式の最小値がその解となる

\*正則化パラメータ $\lambda = \alpha/\beta$



# 1.2.6ベイズ曲線フィッティング

完全なベイズアプローチ

加法・乗法定理を矛盾なく適用し、 $w$ のすべての値に関して積分する

→このような周辺化はパターン認識のベイズ手法の根幹

予測分布 $p(t|x, x, t)$ を評価する

$$p(t|x, x, t) = \int p(t|x, w)p(w|x, t)dw$$

上式の事後分布はガウス分布となる

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

平均  $m(x) = \beta\phi(x)^T S \sum_{n=1}^N \phi(x_n)t_n$

分散  $s^2(x) = \frac{\beta^{-1}}{\phi(x)^T S \phi(x)}$

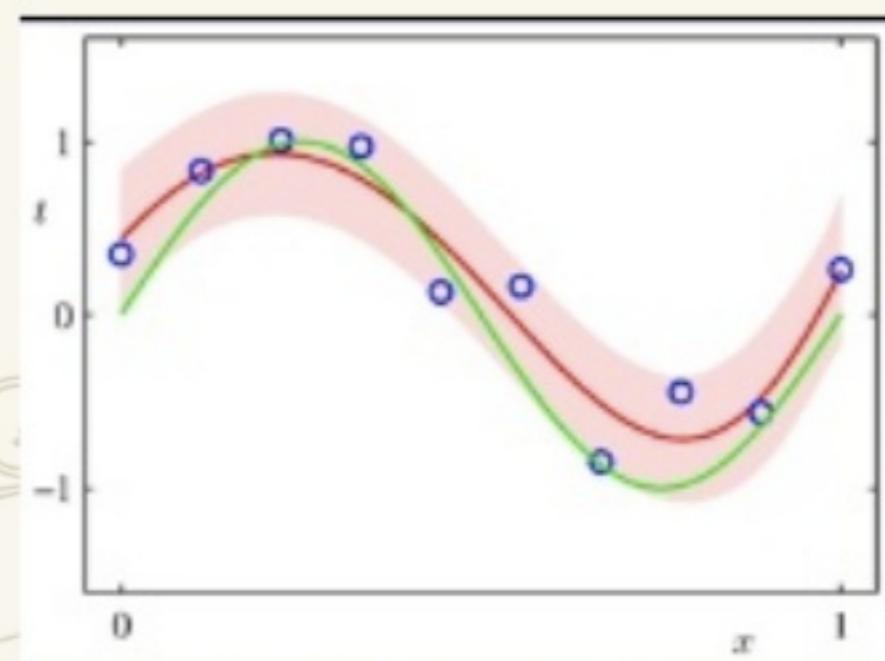
$t$ の予測値の目標変数による不確実性

パラメータ $w$ に対する不確実性（ベイズ的）

M=9次多項式のベイズフィッティング

赤線：予測分布の平均

赤い領域：平均 $\pm 1$ 標準偏差範囲



# 1.3 モデル選択

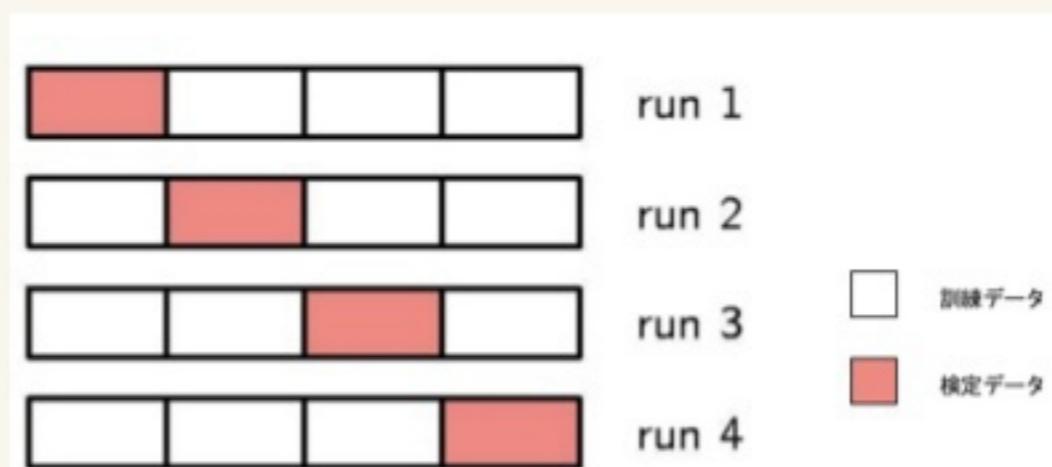
複雑なモデルを制御するパラメータを決めたい

(新たなデータに対してもいい予測ができる)

ホールドアウトの話は前述した通り。

交差確認 (cross validation) を用いる

得られたデータのうち、 $(S-I)/S$ の割合を訓練データとし、全データを評価  
データが少ない時は、LOO法(leave-one-out method)を使う



モデルの複雑さを測る情報量規準

AIC (赤池情報量規準)

$$\ln p(\mathcal{D}|\mathbf{w}_{ML}) - M$$

左式が最大化となるものを選択  
(誤差関数最小化の方が一般)

BIC (ベイズ情報量規準)

4.4.1節で議論のこと

# 情報量規準と言えば

東工大の渡辺先生がWAIC、WBICという指標を考案し、一部界隈で話題に！

## 広く使える情報量規準(WAIC)

東京工業大学 渡辺澄夫

要約：事後分布が正規分布で近似できないときでも AIC, TIC, DIC が使えないときでも、いつでも汎化誤差を推定できる方法を作りました。

---

$X_1, X_2, \dots, X_n$  : 真の分布  $q(x)$  に従う独立な確率変数

$p(x|w)$  : モデル: パラメータ  $w$  を持つ  $x$  の確率分布

$\varphi(w)$  : パラメータ  $w$  の事前分布

---

$$E_w[\quad] = \frac{1}{Z} \int (\quad) \prod_{i=1}^n p(X_i|w) \varphi(w) dw : \text{事後分布による平均}$$

$p^*(x) = E_w[p(x|w)]$  : 予測分布

---

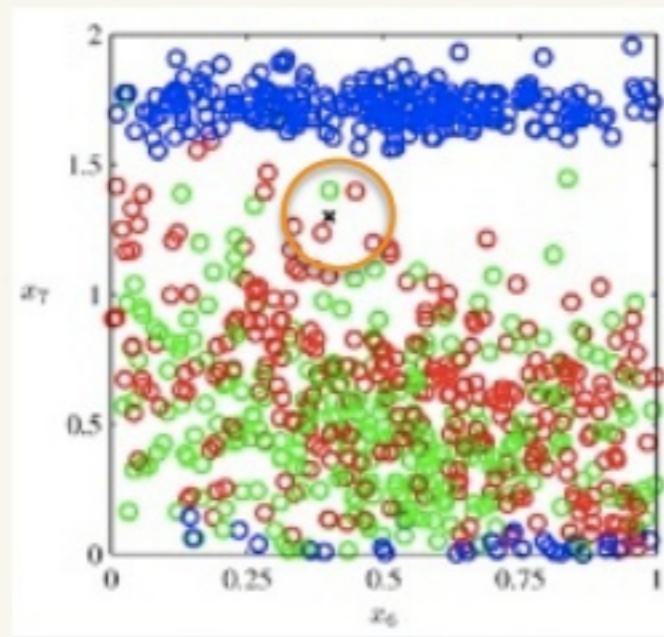
$$T = - (1/n) \sum_{i=1}^n \log p^*(X_i) : \text{学習損失}$$

$$G = - \int q(x) \log p^*(x) dx : \text{汎化損失}$$

$$V = \sum_{i=1}^n \{ E_w[ (\log p(X_i|w))^2 ] - E_w[\log p(X_i|w)]^2 \} : \text{汎関数分散}$$

# 1.4次元の呪い

パターン認識の応用では高次元空間を扱う必要がある



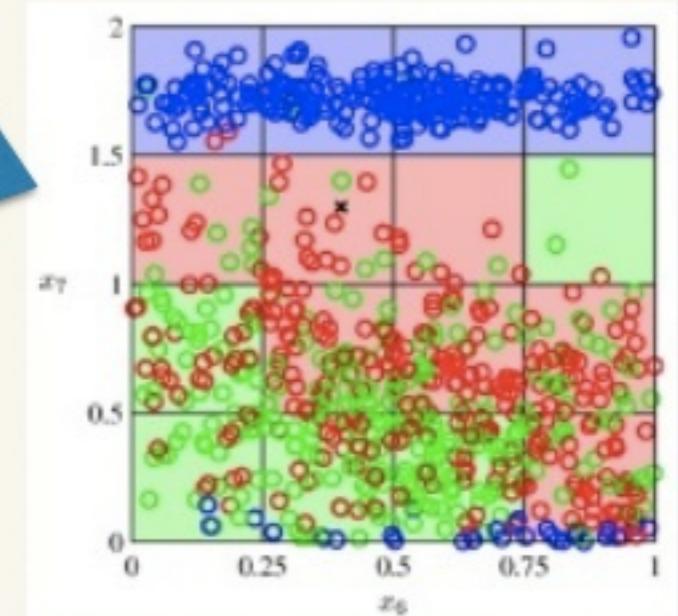
左図xをどの色に分類するか問題

学習アルゴリズムにするには？

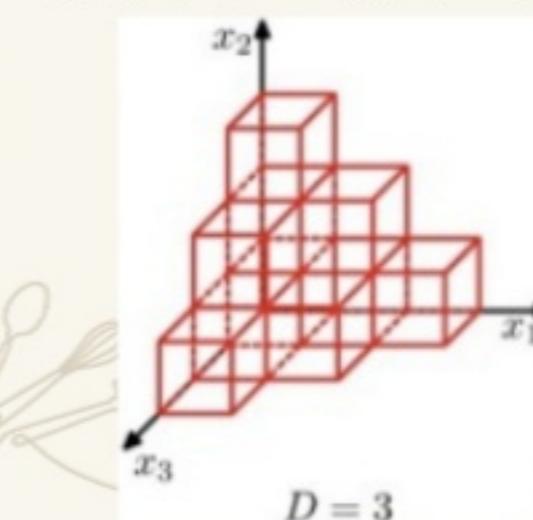
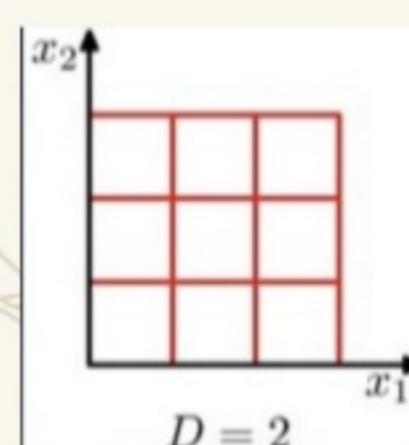
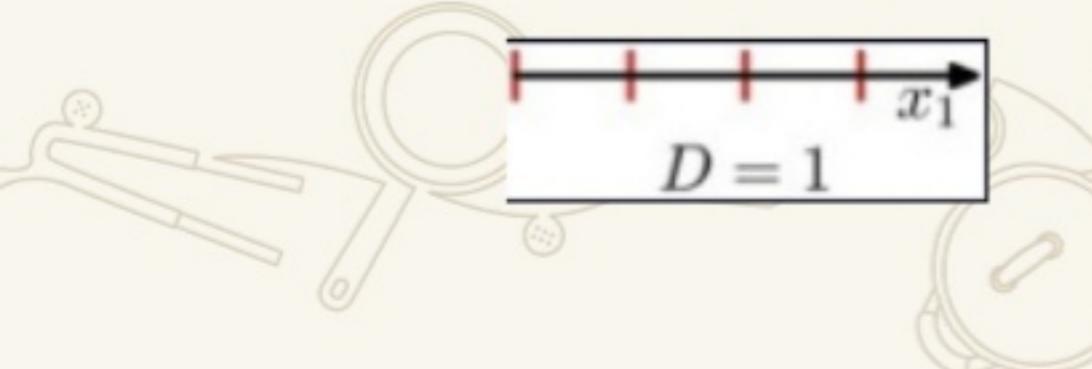
→マス目に区切ってみる



同じマス目に多く含有する  
色に分類する



入力空間が高次元に拡張するとマス目が指数的に増大し、つらい…



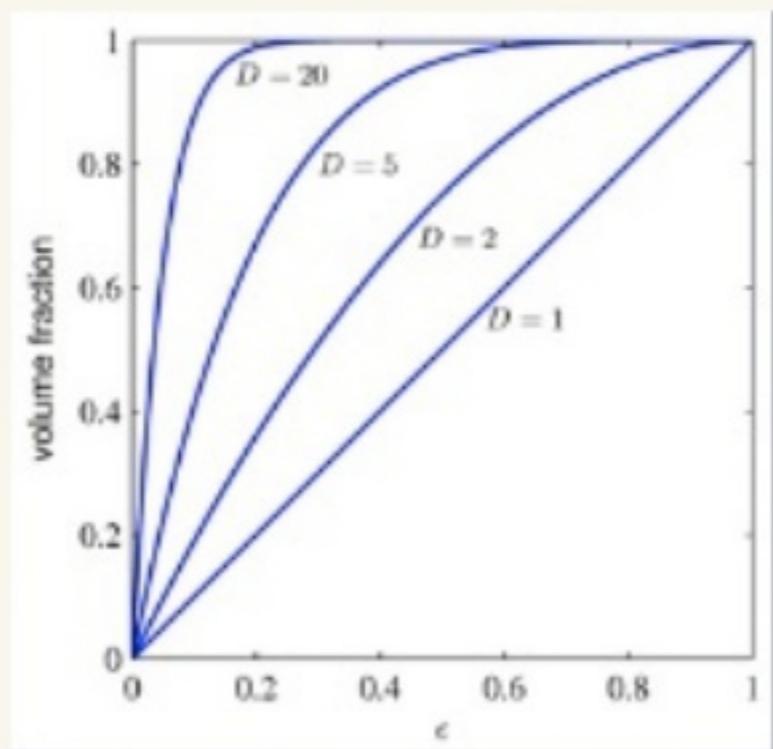
etc..

# 1.4次元の呪い

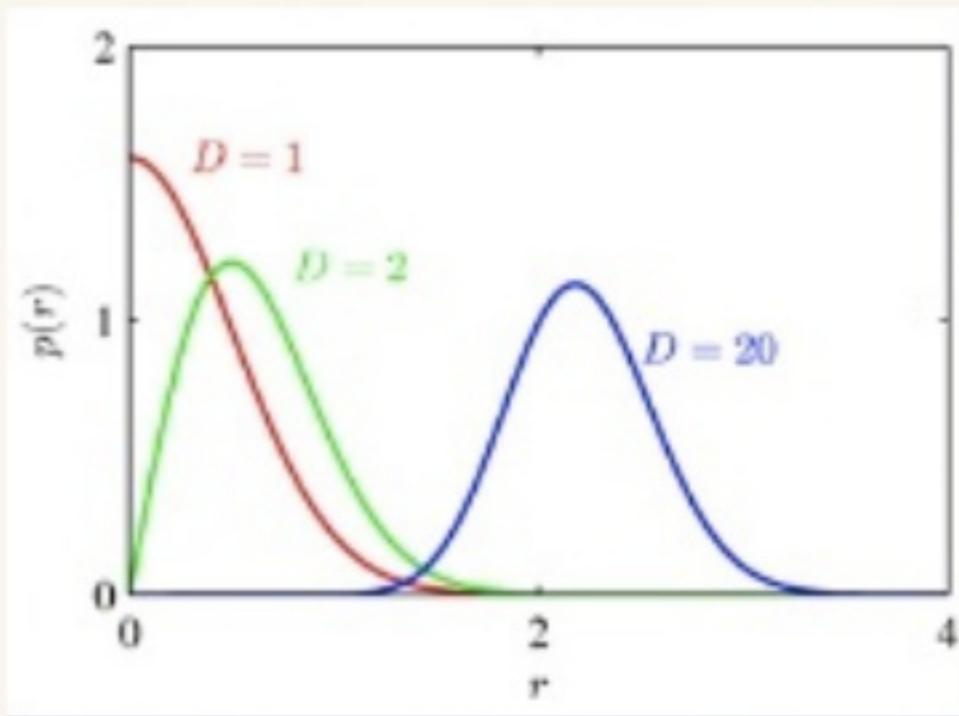
$D$ が増えると、独立な係数の数は $D^3$ に比例する

$M$ 次の多項式では、係数は $D^M$ に増える → 幕乗でヤバい！

3次元空間で生活している我々には高次元空間の幾何的直感は間違う



半径 $r=1-\varepsilon$ と $r=1$ の間にある球の  
様々な次元 $D$ に関するプロット



ガウス分布の半径 $r$ に関する確率密度  
の様々な次元 $D$ に関するプロット

次元の呪い：大きい次元の空間に伴う困難のこと

# 1.4次元の呪い

重要な問題だが、有効な手段がないこともない

- 1、実データは低次元領域であり、目標変数の変化の方向性は限定性が高い
- 2、実データが一般的に滑らかな性質を持っており、局所的に小さな変化しか与えない。→局所的内挿などで対処可能

物体には3つの自由度がある

画像の集合は高次元空間に埋め込まれた3次元の多様体上にある  
→物体の位置、方向、ピクセル強度は複雑なので、高度に非線形  
(パラメータを固定化できれば、少ない数で制御可能になる)

