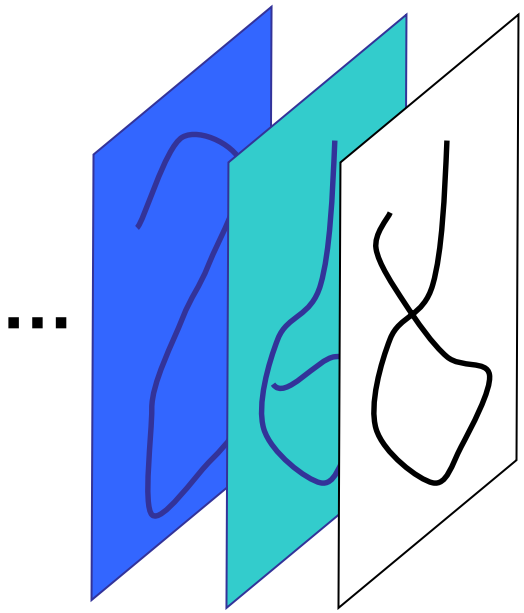
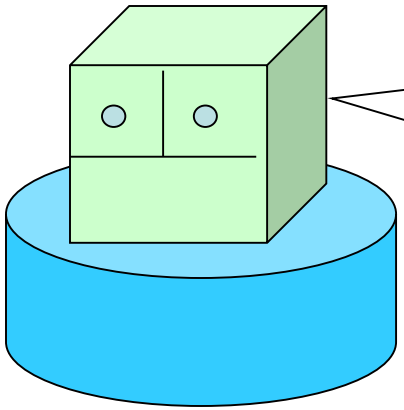
情報学習理論

渡辺澄夫
東京工業大学

教師つき学習

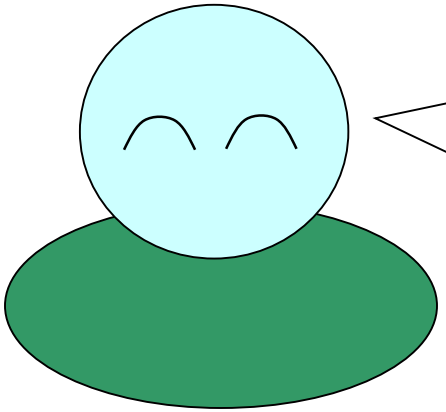


文字の例



先生

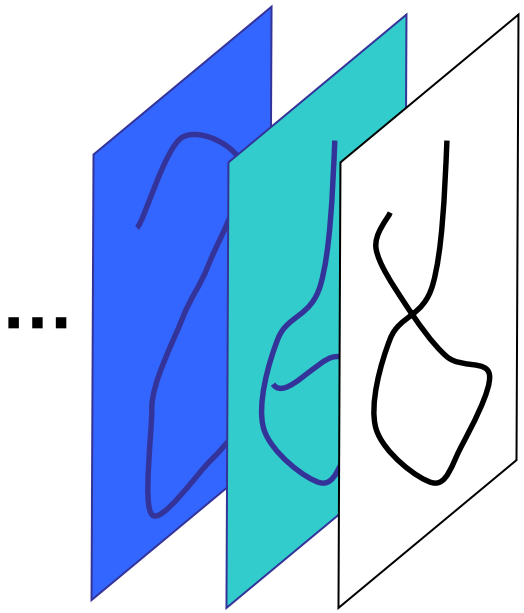
8, 6, 2...
でしょう



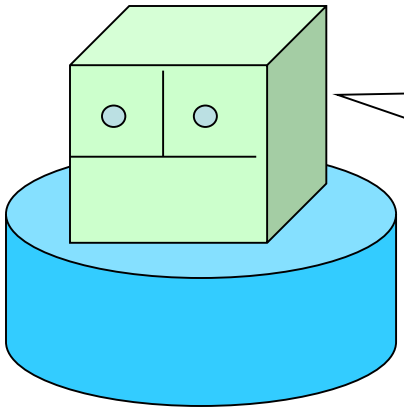
生徒

文字を
読みます

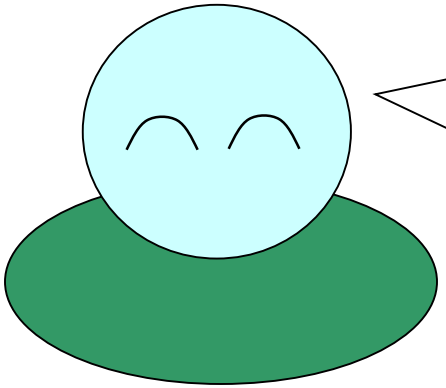
教師つき学習



$X_1, X_2, \ldots, X_n$

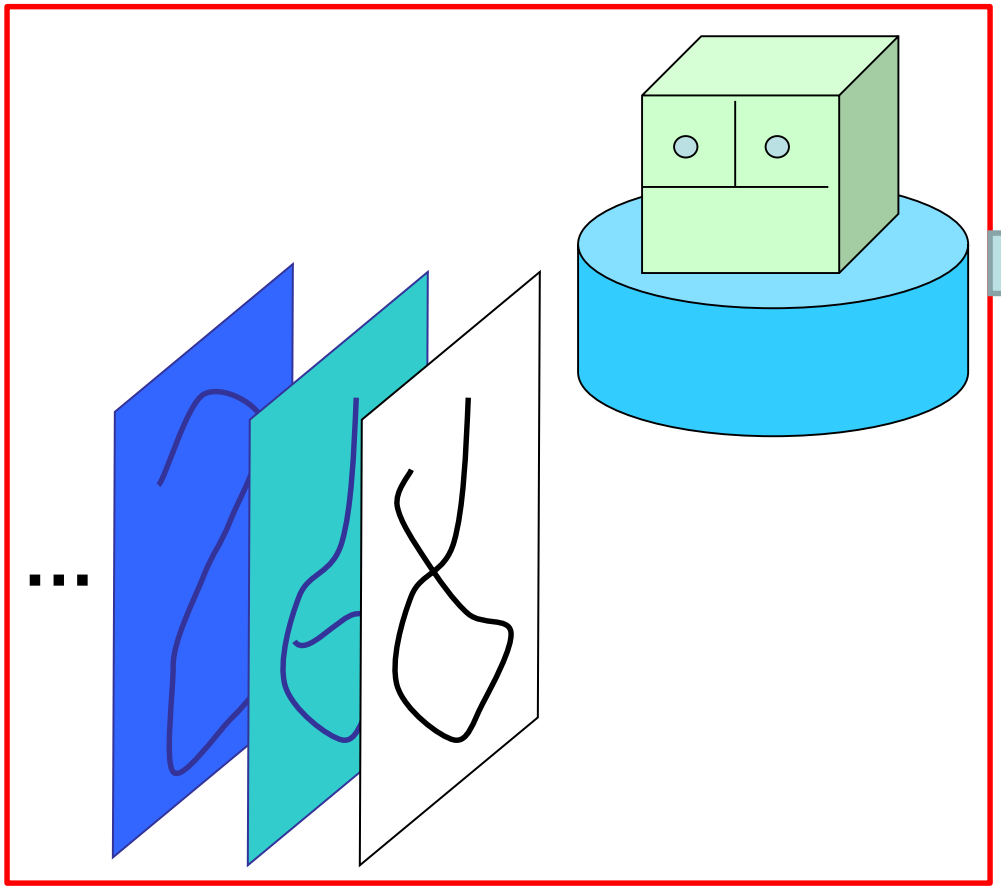


$Y_1, Y_2, \ldots, Y_n$



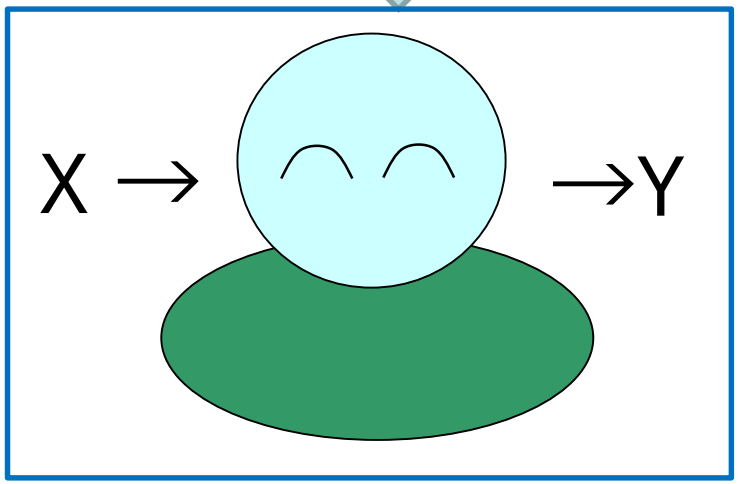
X に対する Y は？

教師つき学習



$X_1, X_2, \ldots, X_n$

$Y_1, Y_2, \ldots, Y_n$



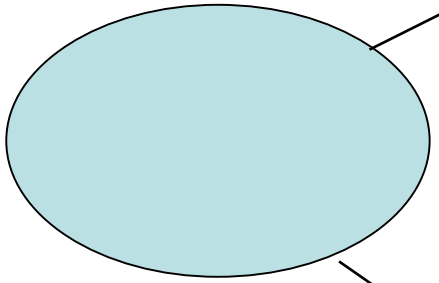
まとめて「情報源」

学習の枠組み

学習データ

$X_1, X_2, \ldots, X_n$

$Y_1, Y_2, \ldots, Y_n$



テストデータ

X

Y

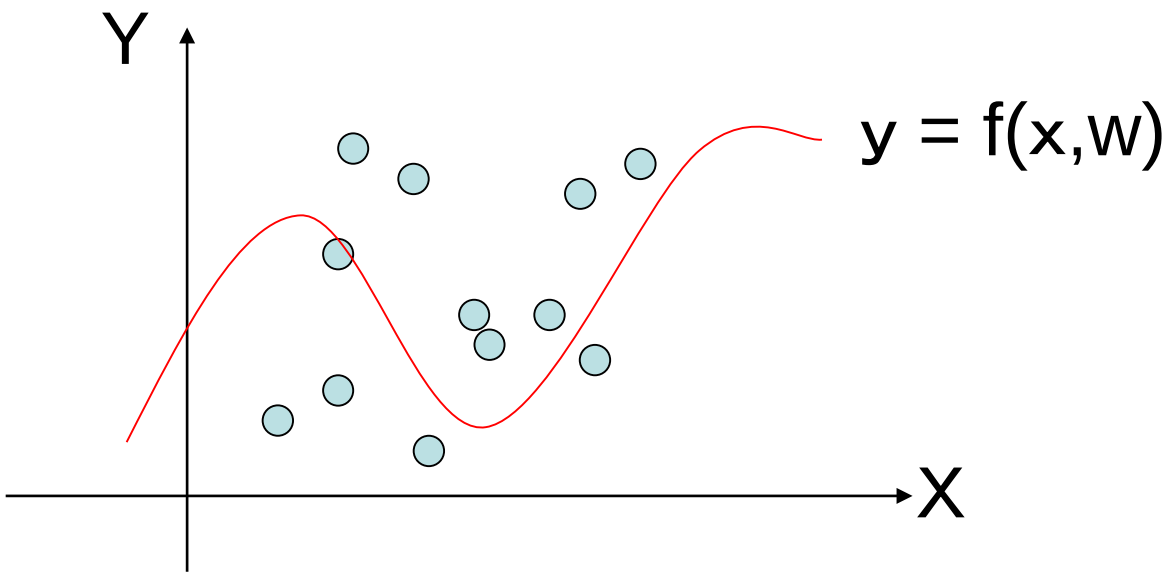
真の情報源

$q(x,y) = q(y|x)\, q(x)$

関数の学習

x は N 次元、y は1次元、w は d次元とする。

サンプルを用いて「x から y への関数」を学習したい。
関数 y = f(x,w) を考える。



二乗誤差

学習誤差関数

$$E(w) = \sum_{i=1}^{n} (Y_i - f(X_i, w))^2$$

E(w) を最小にする w* を求めたい。

◎ 遠い昔 w* を見つけると夢がかなうと信じられていた
◎ 実際は w* は予測には適していないことが分かった。

とはいえ、今日は E(w) の最小化の問題を考察する。

何を探して旅にでるのだろう

前回と今日の違い

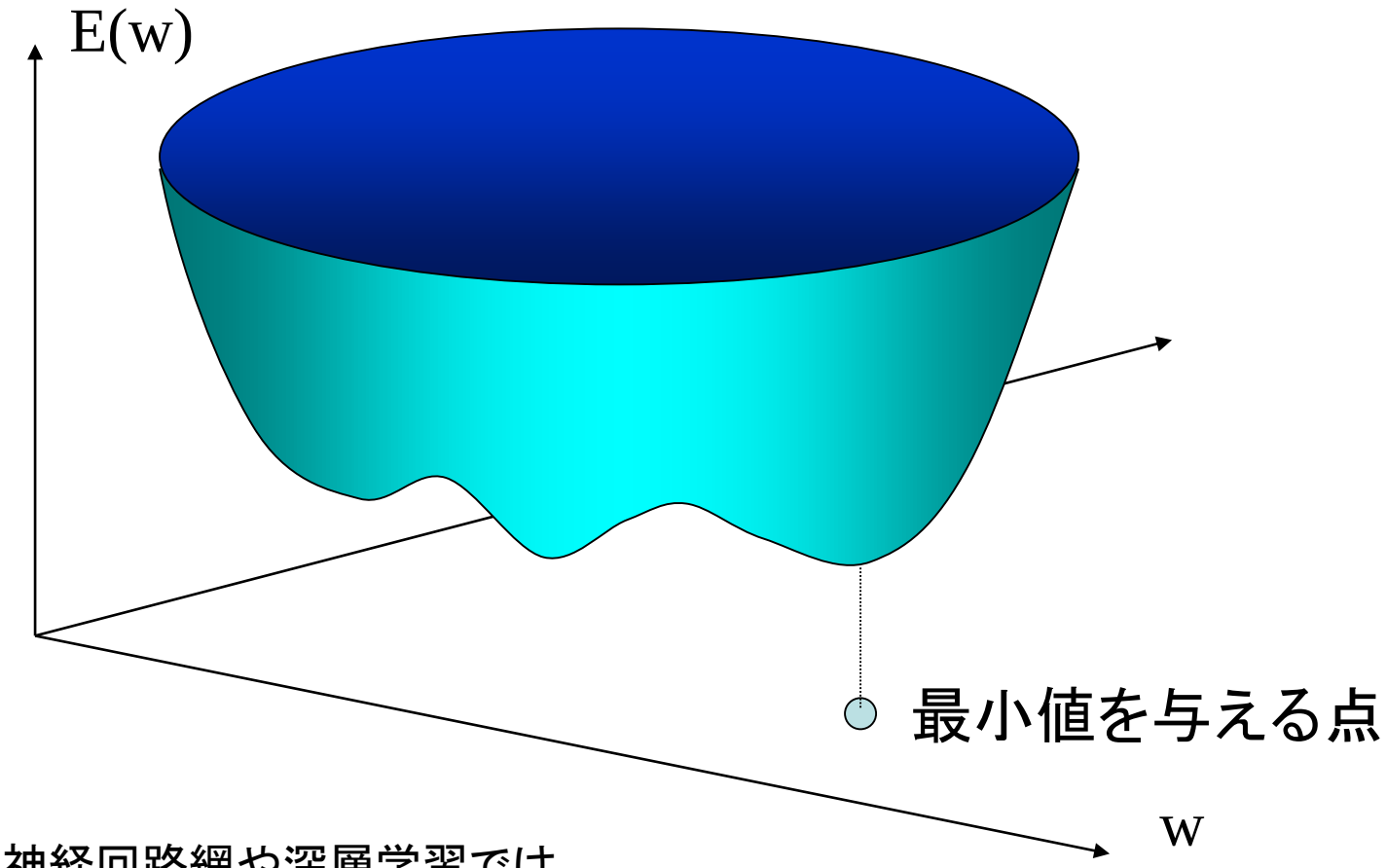
学習誤差関数

$$E(w) = \sum_{i=1}^{n} (Y_i - f(X_i, w))^2$$

において、E(w) が w の2次式ならば、E(w)を
最小にする w* は、∇E=0を解いて
直接求めることができる。

しかし、一般にはそうではない。

誤差曲面



（注） 神経回路網や深層学習では
最小値を与える点は無限遠にあることが多い。
極小付近も最小値付近も2次関数では近似できない。

最適化法

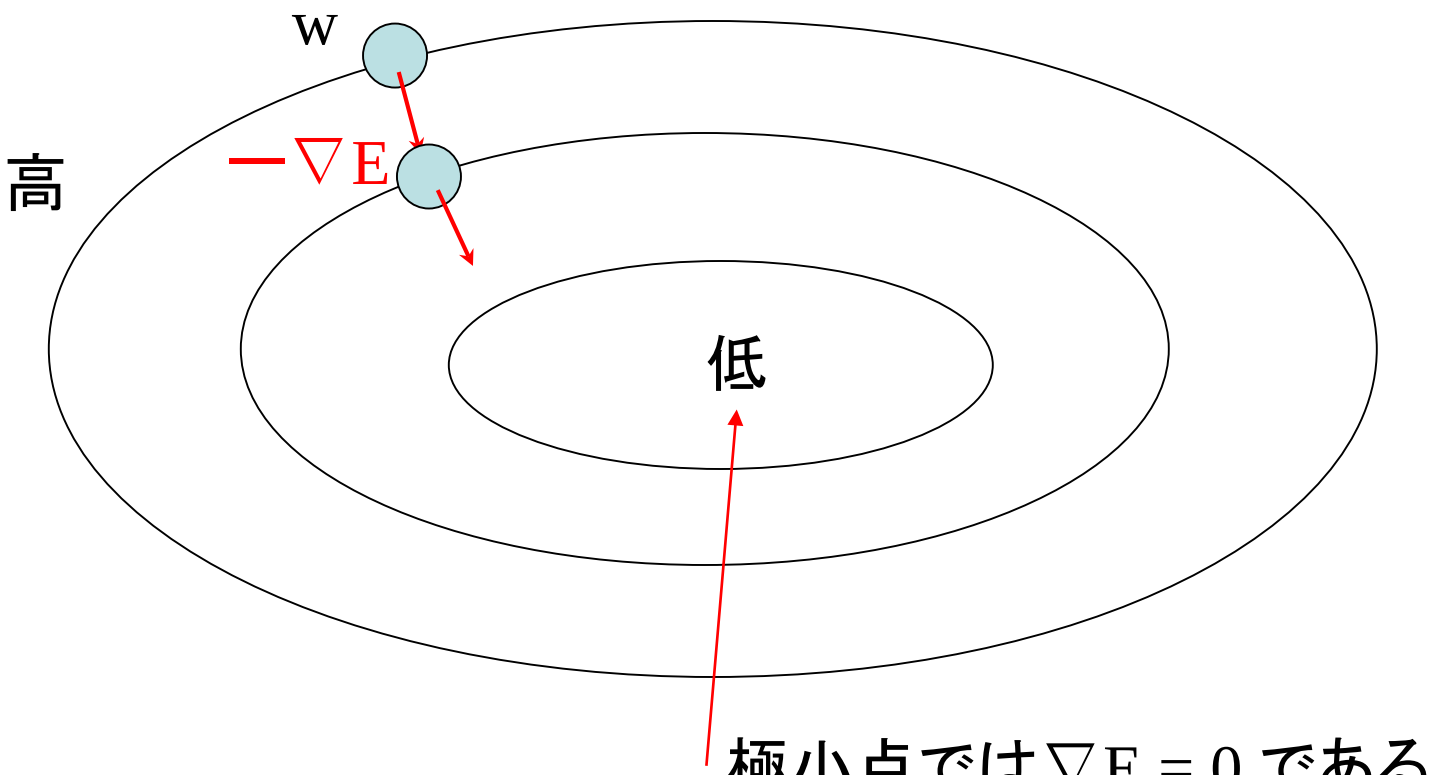
$E(w) = E(w_1, w_2, \ldots, w_d)$ を最小にする w^* を探す

1. 微分方程式 （最急降下法など）
2. 確率アルゴリズム （MCMC・進化計算など）
3. 解くアルゴリズム （非線形計画法など）

復習

$$\nabla E(w) = \left(\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \cdots\cdots, \frac{\partial E}{\partial w_d} \right)$$

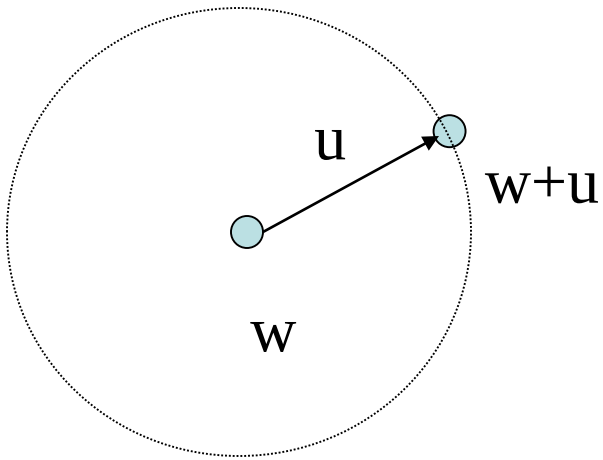
$\nabla E = \mathrm{grad}\, E$ とも書く



なぜ∇E

テーラー展開 : ||u|| が十分に小さいとき

$$E(w+u) = E(w) + u \cdot \nabla E(w) + O(u^2)$$



|| u || = 一定のとき

E(w+u) が最小になるのは

$u \propto -\nabla E(w)$ のとき

（注意）$-\nabla E$ は等高線と直交している

最急降下法

常微分方程式

$$\frac{dw}{dt} = -\nabla E(w)$$

離散化　t=0,1,2,3,…　　η>0 は定数

$$w(t+1) - w(t) = -\eta \nabla E(w(t))$$

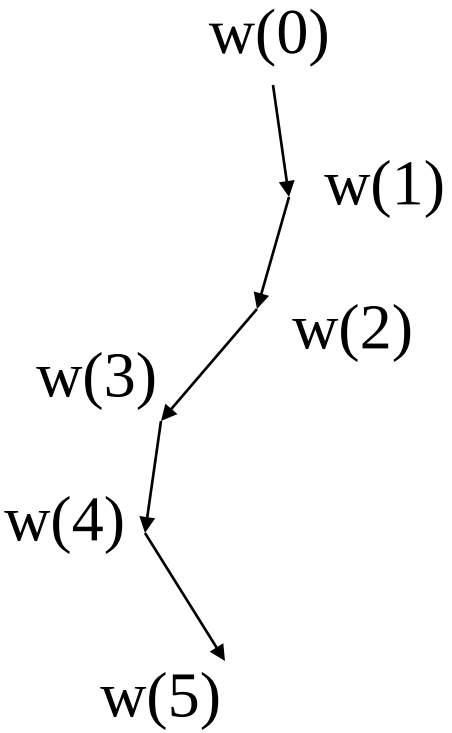
最急降下法

$\eta>0$　：十分小さな定数

(1) 出発点 w(0) をランダムに初期化

(2) $w(n+1) = w(n) - \eta \nabla E(w(n))$

(3) n=0,1,2,…を繰り返す



E(w(n)) はだんだん小さくなってゆく

最急降下法の例

例題　$E(x,y) = x^2/4 + y^2$

最急降下　$-\nabla E = -(x/2, 2y)$

到達点　$\nabla E=0 \Leftrightarrow (x,y)=(0,0)$

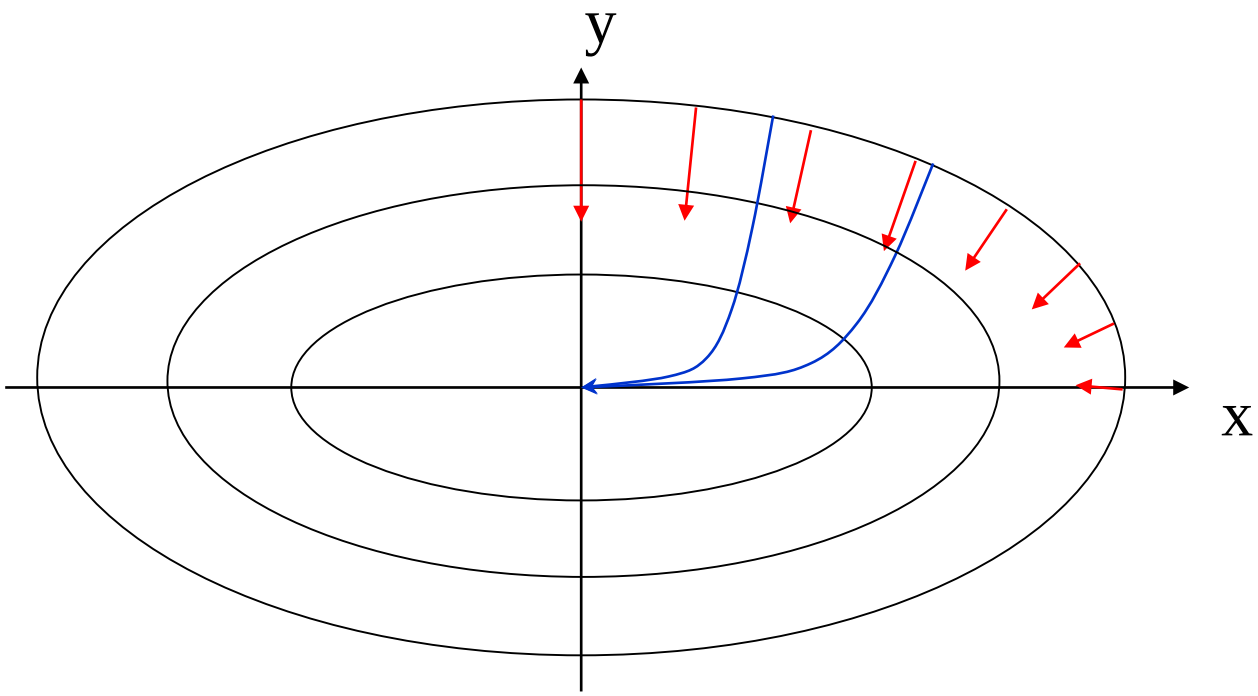
微分方程式

$dx/dt = -x/2$

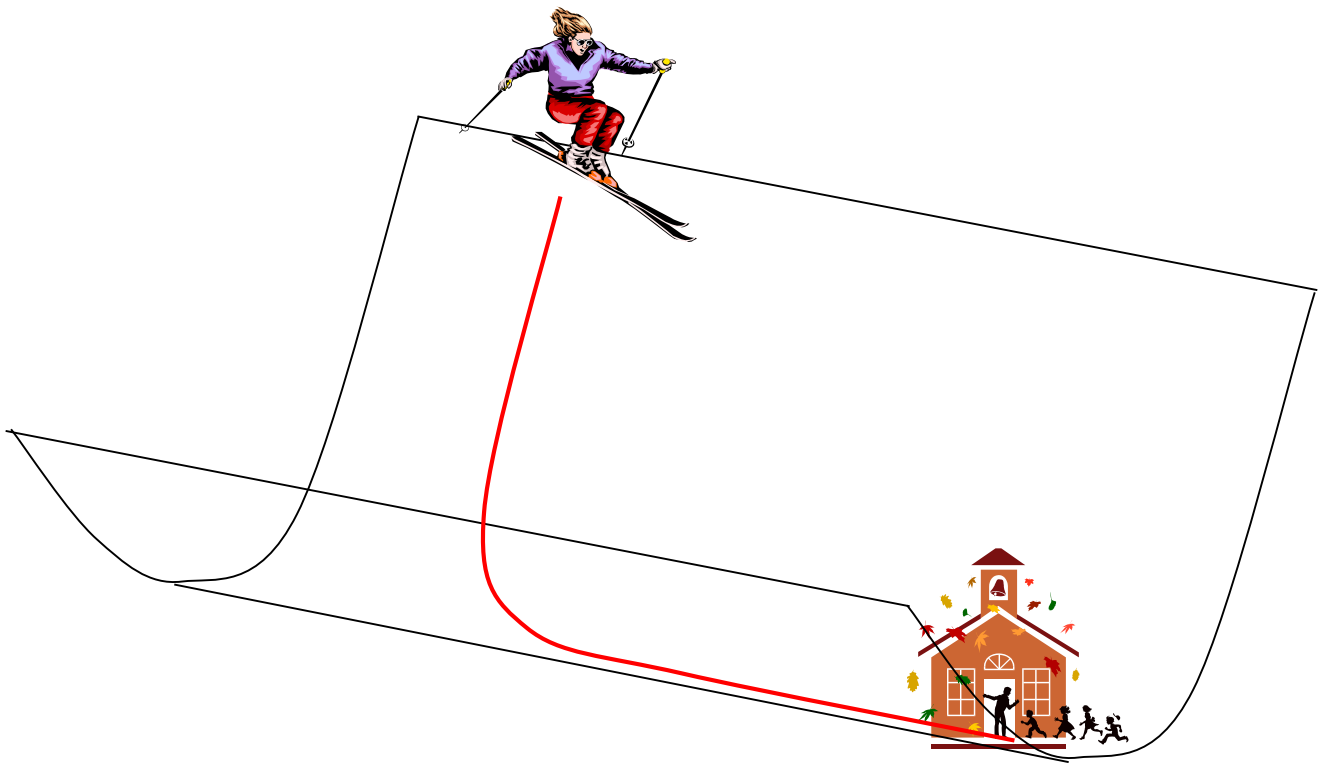
$dy/dt = -2y$

∇E は等高線と直交

→　軌跡は等高線と直交する



直滑降は一番早い？



最急降下法は局所的に見れば $E(w)$ を一番小さくする方向に移動するが、大局的には効率のよい方法ではない。

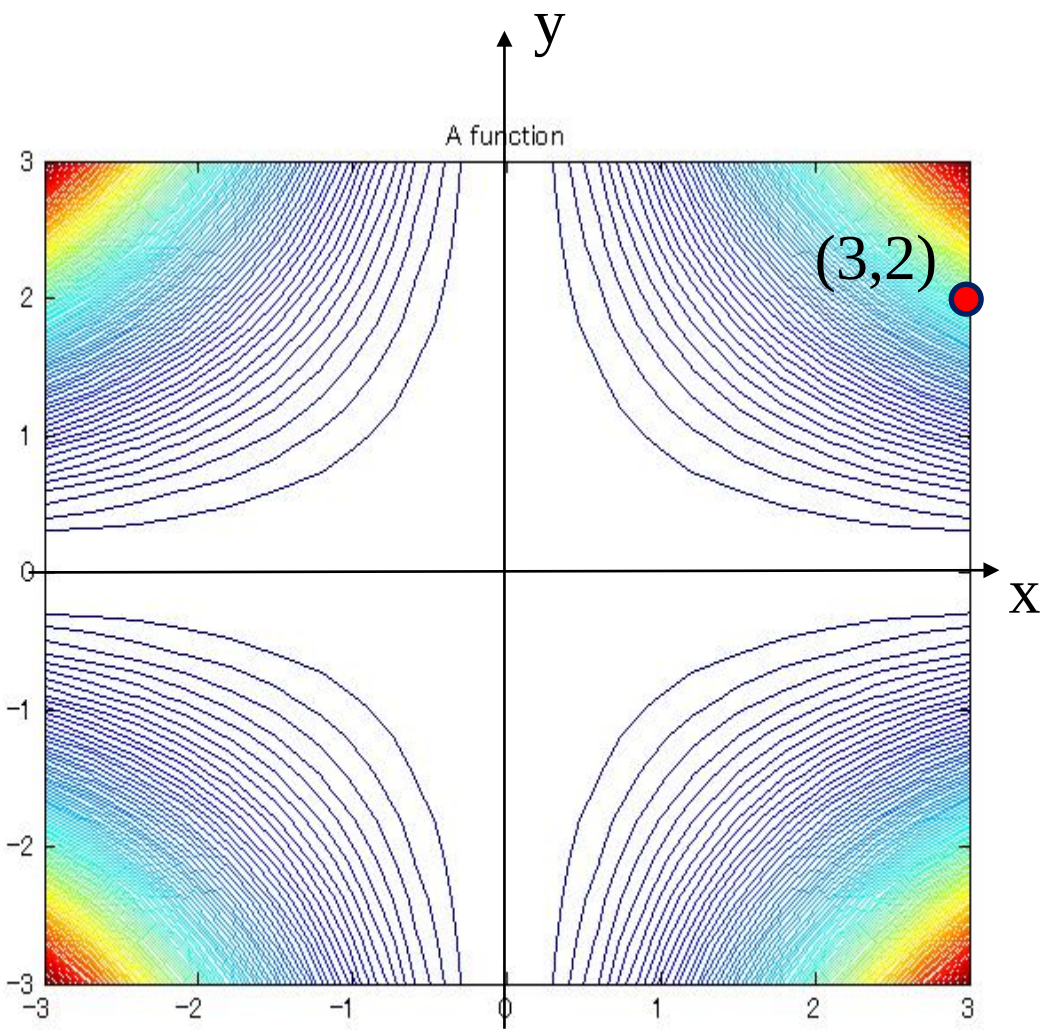
問1　最急降下法

$E(x,y) = x^2 y^2$

(1) $\nabla E=0$となる
　　集合を求めよ

(2) 等高線をかけ

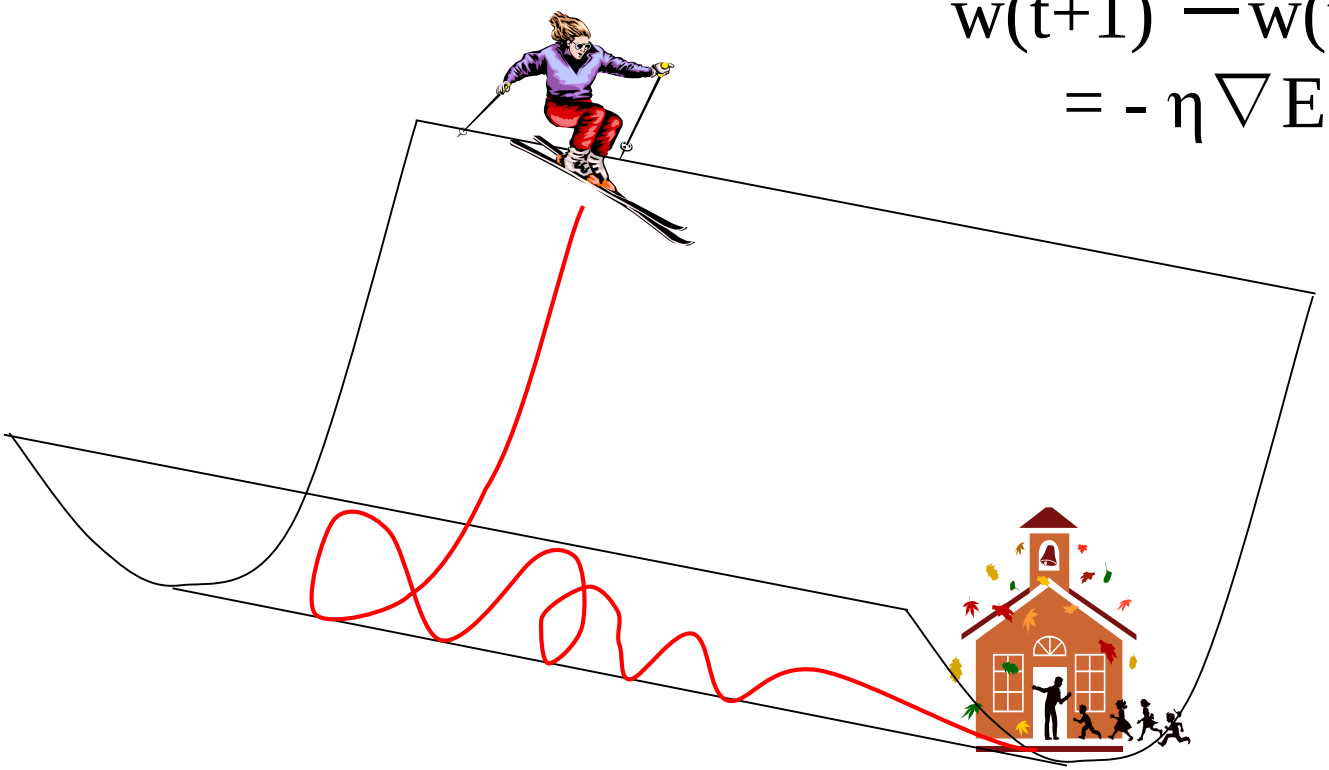
(3) 点(3,2) を
　　初期値とする
　　最急降下法の
　　軌跡をかけ



山の向こうまで探しに行く

離散化の影響

$$w(t+1) - w(t) = -\eta \nabla E(w(t))$$

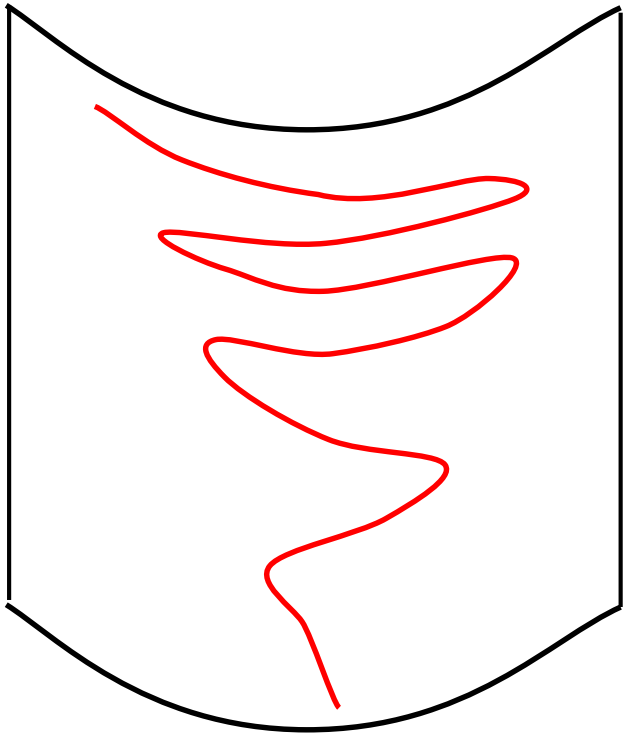


定数 $\eta>0$ が小さすぎると進みかたが遅くなる
大きすぎると振動が起こりやすくなる。

慣性項の追加

離散化　t=0,1,2,3,…　　$\eta>0$, $\alpha>0$ は定数

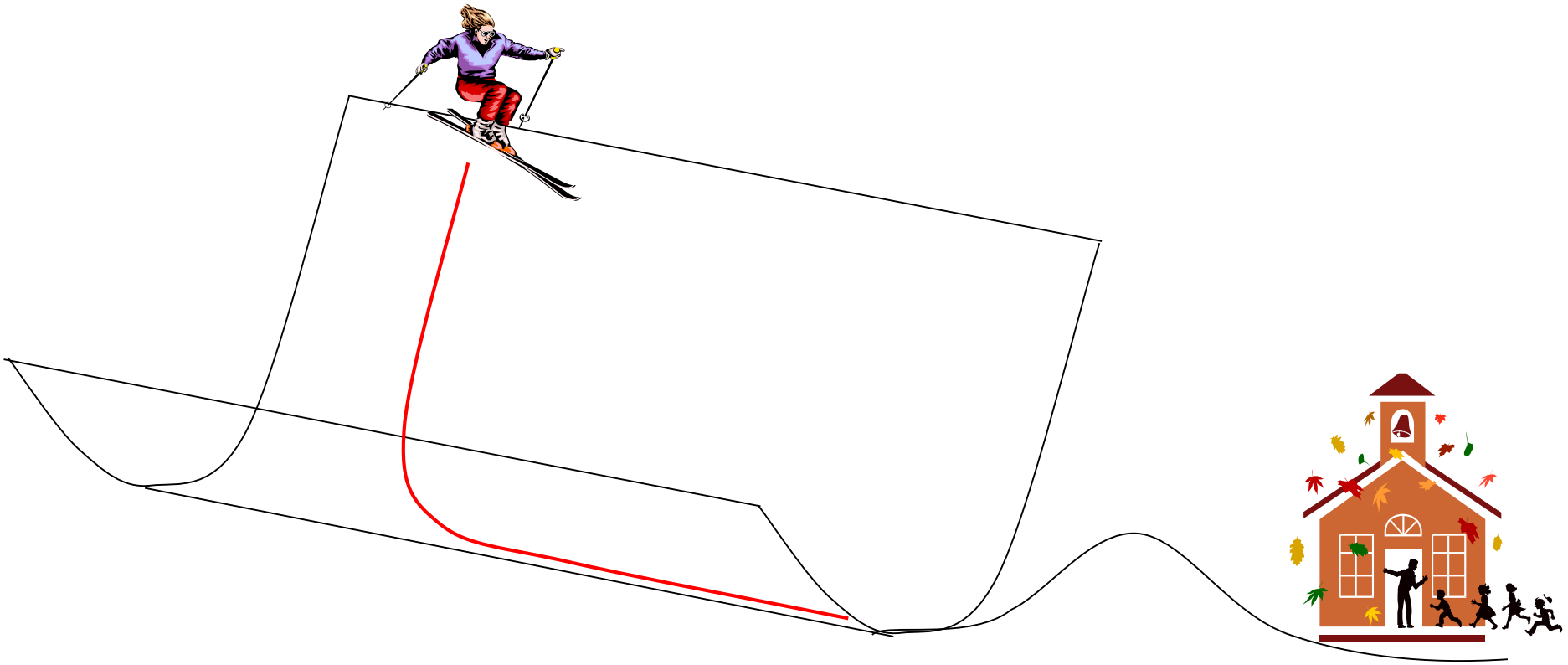
$$w(t+1) - w(t) = -\eta \nabla E(w(t)) + \alpha (w(t) - w(t-1))$$



直線的に進むときはどんどん加速。

小さな振動部分は打ち消される。
大きな振動部分は増幅される。

直滑降で辿りつけるか？



最急降下法は局所解で止まる

確率項の追加

離散化　$t=0,1,2,3,\ldots$　　$\eta>0$ は定数

$$w(t+1) - w(t) = -\eta \nabla E(w(t)) + \text{正規乱数}$$

（注意）

◎　乱数を含めることで局所解を確率的に脱出できる。
　　乱数が小さいと局所解を抜け出る確率が小さく
　　乱数が大きいと最適点での収束が遅い。
　　これは最適化問題が持つ一般的な構造。

◎　「最適解に近づいたら乱数を小さくする」とよいように思えるが
　　最適解か局所解かを判定することは容易にできることではない。

◎　確率１で大局解をみつけることができる乱数の制御法は理論的には
　　分かっているが、演算量が多大すぎて今のところ使えない。

◎　そもそも「どんな最適化問題でも効率的に解ける」という方法はない。
　　また「最適解がみつかったこと」を確認する方法もない。

伊藤清の確率微分方程式

離散化して乱数を含めたものを連続に戻してみる

$$\frac{dw}{dt} = -\nabla E(w) + \frac{d(\text{正規乱数})}{dt}$$

（注意）

◎「正規乱数の微分」は超関数の値をとる確率変数（伊藤清）。

◎ 物理学では この方程式を Langevin 方程式という。
確率微分方程式の一種で実はこれは解ける。

◎ 解の密度関数は Fokker-Planck 方程式に従う。
その解は、伊藤清の確率積分＝Feynman の経路積分。

◎ 解の $t \to \infty$ は平衡状態に収束する。正規乱数の分散が $t\sigma^2$ のとき、収束先の確率分布は $\propto \exp(- (1/2\sigma^2) E(w))$ 。
$\sigma^2 = V[Y-f(X,w)]$ のときベイズ推測と等価になる。

(注意） 学習って何だっけ

学習誤差関数 $E(w) = \sum_{i=1}^{n} (Y_i - f(X_i, w))^2$

E(w) を最小にする w* は汎化誤差を最小にしない。最小化問題を解くことは予測精度の点からは積極的な意味を有していない。

学習を途中で止めたり、exp(- E(w)) からの w で平均するほうが予測誤差は小さくなる。

最急降下法は「予測がうまくいくように途中で止める」ために利用されているだけで最小点をみつけるために使われているのではない・・・。

探していたものは本当の◎◎ではなかった。それは探していた途中にあったのだ。しかし見つけていたことに気づけるのは後になってから・・・。

問2　最急降下法

最急降下法について次の3つのパラメータ設計を考える。

最急降下の一回の大きさ　$\eta>0$

加速度項の大きさ　　$0< \alpha <1$

正規乱数の大きさ

人工的に作った評価関数 $E(x,y) = (5\sin(x)-y)^2 + 0.02\, x^4 – 2x^2$
の最小点を最急降下法で探す問題において上記のパラメータがどのような影響を与えるかを実験的に調べてみよう。

$E(x,y) = E(-x,-y)$ が成り立つので答えはひとつではない。

問2　最急降下法

	小さい	大きい
定数 η		
慣性項 α		
乱数の大きさ		