

Parameter Optimization Algorithms

Toru Wakahara

[1] 2乗和エラー関数の最小化の意味

[2] エラー関数の最小値探索

- 局所的2次近似

[3] 最小値探索アルゴリズム

- Gradient descent 法

- Line search 法

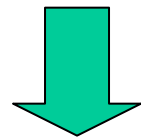
- Conjugate gradients 法

- Newton 法

- Quasi-Newton 法

2乗和エラー関数の最小化の意味(1)

$$\text{2乗和エラー関数: } E = \frac{1}{2} \sum_{n=1}^N \{ y(\mathbf{x}_n; \mathbf{w}) - t^n \}^2$$

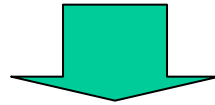


$N \rightarrow \infty$ での振る舞い

$$\begin{aligned} E &= \lim_{N \rightarrow \infty} \frac{1}{2N} \sum_{n=1}^N \{ y(\mathbf{x}_n; \mathbf{w}) - t^n \}^2 \\ &= \frac{1}{2} \iint \{ y(\mathbf{x}; \mathbf{w}) - t \}^2 p(t, \mathbf{x}) dt d\mathbf{x} \\ &= \frac{1}{2} \iint \{ y(\mathbf{x}; \mathbf{w}) - t \}^2 p(t | \mathbf{x}) p(\mathbf{x}) dt d\mathbf{x} \end{aligned}$$

2乗和エラー関数の最小化の意味(2)

$$E = \frac{1}{2} \iint \{y(\mathbf{x}; \mathbf{w}) - t\}^2 p(t | \mathbf{x}) p(\mathbf{x}) dt d\mathbf{x}$$



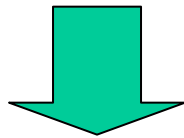
$$\langle t | \mathbf{x} \rangle \equiv \int t p(t | \mathbf{x}) dt, \quad \langle t^2 | \mathbf{x} \rangle \equiv \int t^2 p(t | \mathbf{x}) dt$$

を用いて変形すると

$$\begin{aligned} \{y - t\}^2 &= \{y - \langle t | \mathbf{x} \rangle + \langle t | \mathbf{x} \rangle - t\}^2 \\ &= \{y - \langle t | \mathbf{x} \rangle\}^2 + 2\{y - \langle t | \mathbf{x} \rangle\} \{\langle t | \mathbf{x} \rangle - t\} \\ &\quad + \{\langle t | \mathbf{x} \rangle - t\}^2 \end{aligned}$$

2乗和エラー関数の最小化の意味(3)

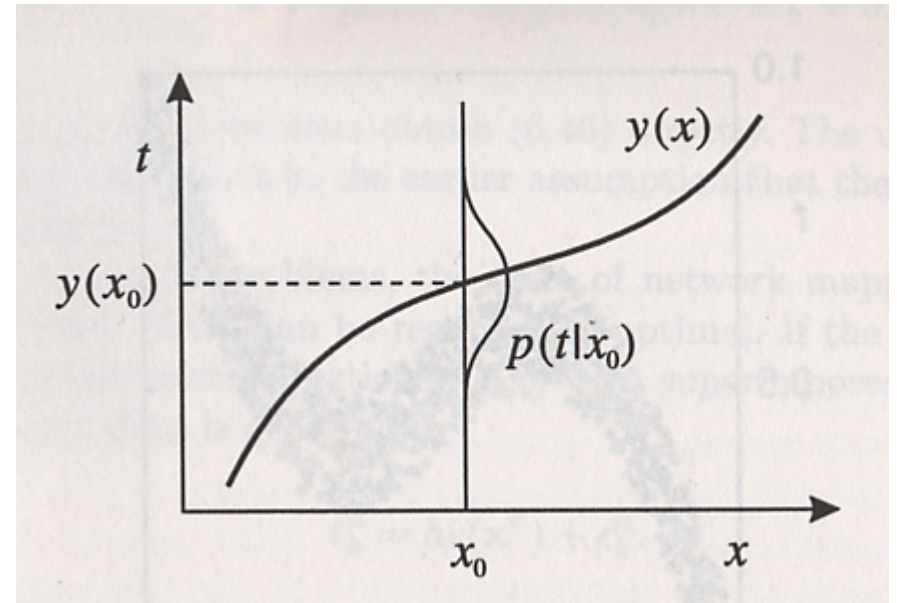
$$E = \frac{1}{2} \int \{ y(\mathbf{x}; \mathbf{w}) - \langle t | \mathbf{x} \rangle \}^2 p(\mathbf{x}) d\mathbf{x} \\ + \frac{1}{2} \int \{ \langle t^2 | \mathbf{x} \rangle - \langle t | \mathbf{x} \rangle^2 \} p(\mathbf{x}) d\mathbf{x}$$



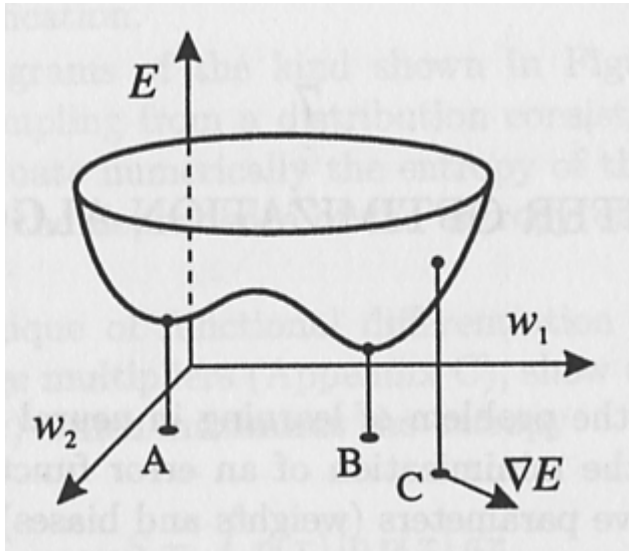
E の最小化:

$$y(\mathbf{x}; \mathbf{w}^*) = \langle t | \mathbf{x} \rangle$$

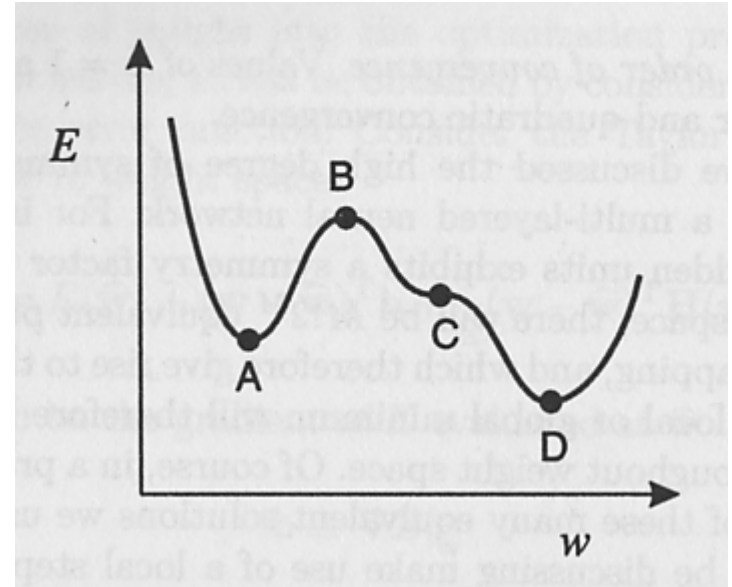
最適解



エラー曲面 $E(w)$



A, B は $E(w)$ の極小点
C での勾配はベクトル ∇E で与えられる



停留点 $\nabla E = 0$ の意味

A: local minimum

B: local maximum

C: saddle point

D: global minimum

$E(\boldsymbol{w})$ の最小値探索

$E(\boldsymbol{w})$ の非線形性により、一般に $\nabla E = 0$ を
閉形式で解くことはできない



重み空間 \boldsymbol{w} の逐次探索により解を求める

$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} + \Delta \boldsymbol{w}^{(\tau)}$$

課題:

- (1) 初期値 $\boldsymbol{w}^{(0)}$ をどう選ぶか
- (2) 更新量 $\Delta \boldsymbol{w}^{(\tau)}$ をどう選ぶか



the *global minimum* に到達できるか

$E(\mathbf{w})$ の局所的2次近似(1)

$E(\mathbf{w})$ を点 $\hat{\mathbf{w}}$ の周りに Taylor 展開すると

$$E(\mathbf{w}) = E(\hat{\mathbf{w}}) + (\mathbf{w} - \hat{\mathbf{w}})^t \mathbf{b} + \frac{1}{2} (\mathbf{w} - \hat{\mathbf{w}})^t \mathbf{H} (\mathbf{w} - \hat{\mathbf{w}})$$

$$\mathbf{b} \equiv \nabla E \Big|_{\hat{\mathbf{w}}}, \quad (\mathbf{H})_{ij} \equiv \frac{\partial E}{\partial w_i \partial w_j} \Big|_{\hat{\mathbf{w}}} : \text{the Hessian matrix}$$

$$\nabla E = \mathbf{b} + \mathbf{H} (\mathbf{w} - \hat{\mathbf{w}})$$

$E(\mathbf{w})$ の a minimum \mathbf{w}^* の近傍を2次近似すると

$$E(\mathbf{w}) = E(\mathbf{w}^*) + \frac{1}{2} (\mathbf{w} - \mathbf{w}^*)^t \mathbf{H} (\mathbf{w} - \mathbf{w}^*)$$

$E(\mathbf{w})$ の局所的2次近似(2)

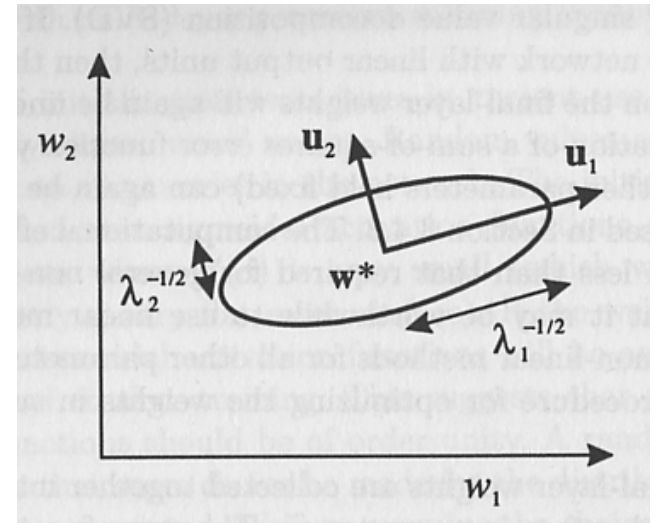
Hessian matrix H の固有値問題を解くと

$$H \mathbf{u}_i = \lambda_i \mathbf{u}_i, \quad \mathbf{u}_i^t \mathbf{u}_j = \delta_{ij}$$

固有ベクトルによる展開式 $\mathbf{w} - \mathbf{w}^* = \sum_i \alpha_i \mathbf{u}_i$ を
 $E(\mathbf{w})$ の式に代入すると

$$E(\mathbf{w}) = E(\mathbf{w}^*) + \frac{1}{2} \sum_i \lambda_i \alpha_i^2$$

$E(\mathbf{w}) = \text{const.}$ は超楕円体



課題1

The Hessian matrix H が正定値 (*positive definite*):

$$\mathbf{v}^t H \mathbf{v} > 0 \quad \text{for all } \mathbf{v} \neq \mathbf{0}$$

であれば, H の固有値がすべて正であることを示しなさい.

次に, H が正定値であることが

$$E(\mathbf{w}) = E(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^t H (\mathbf{w} - \mathbf{w}^*)$$

において \mathbf{w}^* が a unique global minimum であることの必要十分条件であることを証明しなさい.

Gradient descent 法(1)

$$E = \frac{1}{2} \sum_{n=1}^N \{ y(\mathbf{x}_n; \mathbf{w}) - t^n \}^2 \rightarrow \min \text{ for } \mathbf{w}$$

[1] 初期値の設定: \mathbf{w}_0

[2] \mathbf{w} の逐次更新: η (learning rate) の導入

$$\Delta \mathbf{w}^{(\tau)} = -\eta \nabla E \Big|_{\mathbf{w}^{(\tau)}} \quad : \text{バッチ処理型}$$

$$\Delta \mathbf{w}^{(\tau)} = -\eta \nabla E^n \Big|_{\mathbf{w}^{(\tau)}} \quad : \text{逐次処理型}$$

→ 振舞いにどんな違いがあるだろうか？

Gradient descent 法(2)

w^* の近傍での収束性の問題

$$\nabla E = H(w - w^*) = \sum_i \alpha_i \lambda_i u_i$$

$$\Delta w = \Delta(w - w^*) = \sum_i \Delta \alpha_i u_i$$

$\Delta w = -\eta \nabla E|_w$ より次の関係が得られる

$$\Delta \alpha_i = -\eta \lambda_i \alpha_i \quad \rightarrow \quad \alpha_i^{new} = (1 - \eta \lambda_i) \alpha_i^{old}$$

T 回の更新を行うと

$$\alpha_i^{(T)} = (1 - \eta \lambda_i)^T \alpha_i^{(0)}$$

Gradient descent 法(3)

w^* の近傍での収束性の問題(続き)

一方, 固有ベクトルの正規直交性より

$$u_i^t (w - w^*) = \alpha_i$$

すなわち, α_i は u_i 方向の w^* への距離を表す

$$\alpha_i^{(T)} = (1 - \eta\lambda_i)^T \alpha_i^{(0)} \rightarrow 0 \quad \text{as } T \rightarrow \infty$$

if $|1 - \eta\lambda_i| < 1$

→ w^* に収束するための η の条件は?

Gradient descent 法(4)

w^* の近傍での収束性の問題(続き)

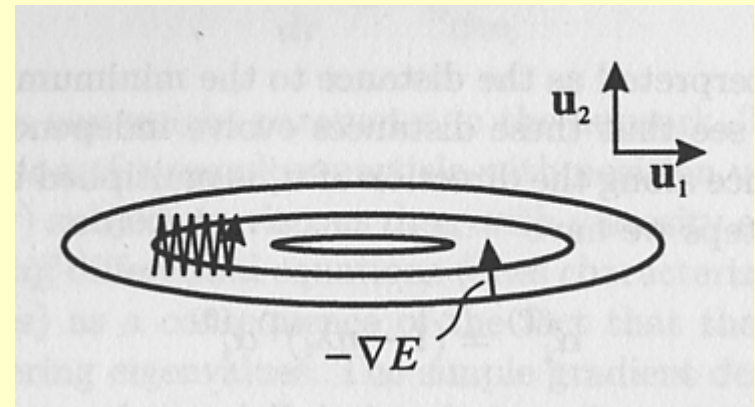
収束条件より

$$|1 - \eta \lambda_{max}| < 1 \rightarrow \eta < 2/\lambda_{max}$$

収束速度は最小の λ_{min} により次式で支配される

$$\left(1 - \frac{2\lambda_{min}}{\lambda_{max}} \right)$$

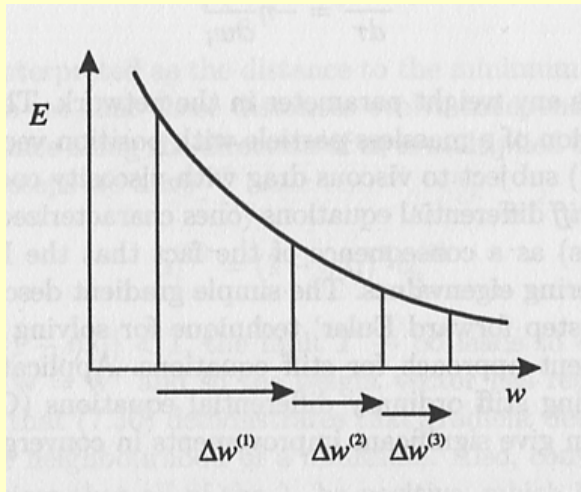
$\lambda_{min} / \lambda_{max}$ が小さいと
収束は極めて遅くなる！



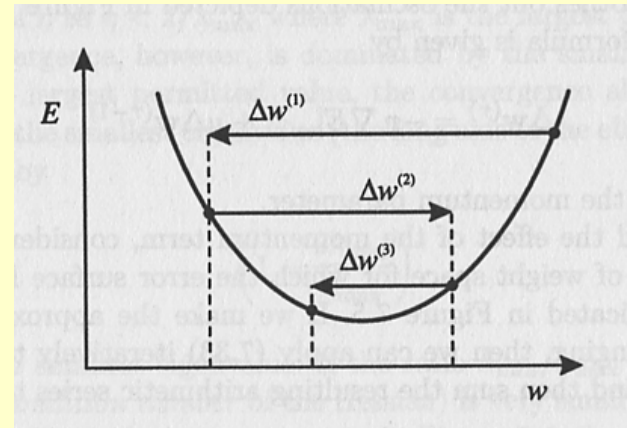
Gradient descent 法(5)

慣性項 *momentum* の導入

$$\Delta \mathbf{w}^{(\tau)} = -\eta \nabla E \Big|_{\mathbf{w}^{(\tau)}} + \mu \Delta \mathbf{w}^{(\tau-1)}$$



曲率が小さいところでは
 $\eta \rightarrow \eta / (1 - \mu)$ により加速



曲率が大きいところでは
慣性項は相殺する
 $\rightarrow \eta$ での更新

課題2

有限幅の刻み $t = \Delta \times \tau$ によるステップ更新の式:

$$\Delta \mathbf{w}^{(\tau)} = -\eta \nabla E \Big|_{\mathbf{w}^{(\tau)}} + \mu \Delta \mathbf{w}^{(\tau-1)}$$

を連続時間での式に移行させると, 次のような運動方程式が得られることを示しなさい.

$$m \frac{d^2 \mathbf{w}}{dt^2} + \nu \frac{d \mathbf{w}}{dt} = -\nabla E$$

$$\text{where } m = \frac{\mu \Delta^2}{\eta}, \quad \nu = \frac{(1 - \mu) \Delta}{\eta}$$

Gradient descent 法(6)

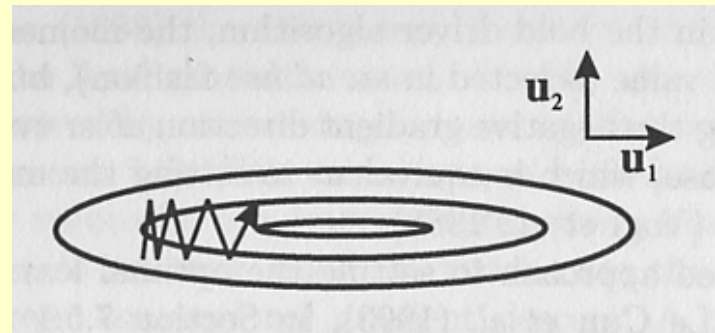
学習率 η の制御法の例:

$$(1) \quad \eta_{new} = \begin{cases} \rho \eta_{old} & \text{if } \Delta E < 0 \\ \sigma \eta_{old} & \text{if } \Delta E > 0 \end{cases}$$

$$\text{ex. } \rho = 1.1, \quad \sigma = 0.5$$

$$(2) \quad \Delta \eta_i^{(\tau)} = \gamma g_i^{(\tau)} g_i^{(\tau-1)}, \quad \gamma > 0$$

$$g_i^{(\tau)} = \frac{\partial E}{\partial w_i^{(\tau)}}$$



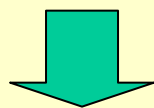
Line search 法(1)

$$E = \frac{1}{2} \sum_{n=1}^N \{ y(\mathbf{x}_n; \mathbf{w}) - t^n \}^2 \rightarrow \text{min for } \mathbf{w}$$

[1] ステップ τ での探索方向の決定: $\mathbf{d}^{(\tau)}$

[2] 方向 $\mathbf{d}^{(\tau)}$ で E が最小になる点を探す

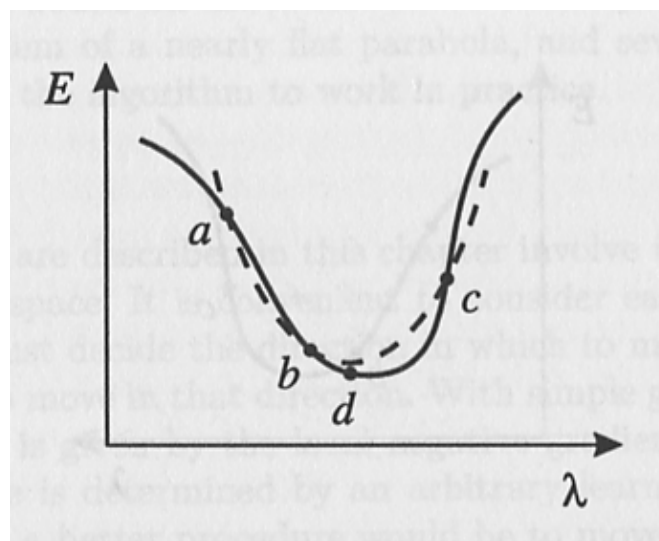
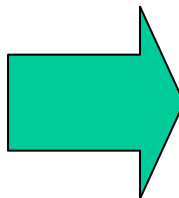
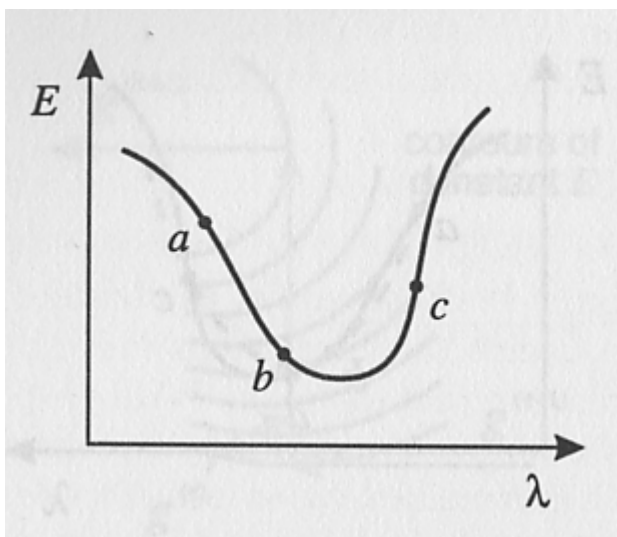
$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \lambda^{(\tau)} \mathbf{d}^{(\tau)}$$



$$E(\lambda) = E(\mathbf{w}^{(\tau)} + \lambda^{(\tau)} \mathbf{d}^{(\tau)}) \rightarrow \text{min for } \lambda^{(\tau)}$$

Line search 法(2)

Parabolic interpolation and Brent's method



最小値を囲い込む

$$a < b < c$$

$$E(a) > E(b) \text{ and } E(c) > E(b)$$

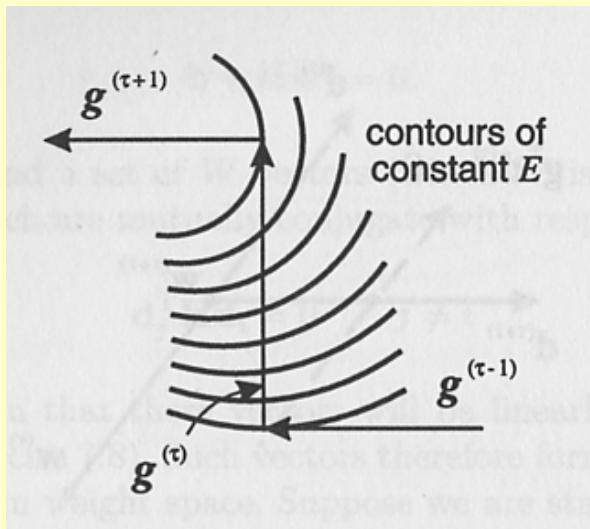
a, b, c を通る放物線で
近似して最小点 d を
求める

→ b, d, c に更新

Conjugate gradients 法(1)

$$E = \frac{1}{2} \sum_{n=1}^N \{ y(\mathbf{x}_n; \mathbf{w}) - t^n \}^2 \rightarrow \min \text{ for } \mathbf{w}$$

Line search では連続する探索方向は直交する



$$\mathbf{g}^{(\tau+1)T} \mathbf{d}^{(\tau)} = 0$$

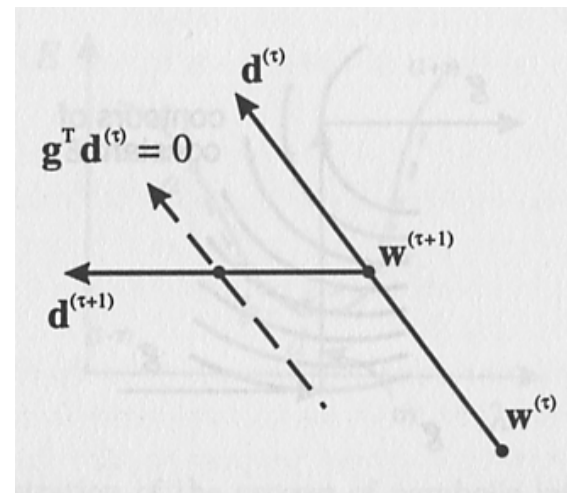
$$\mathbf{g} \equiv \nabla E$$

→ 収束速度が遅い
早くする工夫はないか

Conjugate gradients 法(2)

探索は直交方向に行うが $\mathbf{g} \equiv \nabla E$ を常に監視して
次の直交条件を満たす conjugate vector $\mathbf{d}^{(\tau)}$ を求める

$$\mathbf{g}(\mathbf{w}^{(\tau+1)} + \lambda \mathbf{d}^{(\tau+1)})^t \mathbf{d}^{(\tau)} = 0$$



λ について1次項まで展開すると

$$\mathbf{g}(\mathbf{w}^{(\tau+1)})^t \mathbf{d}^{(\tau)} + \lambda (\mathbf{H} \mathbf{d}^{(\tau+1)})^t \mathbf{d}^{(\tau)} = 0$$

第1項が消えて次式が得られる

$$\mathbf{d}^{(\tau+1)t} \mathbf{H} \mathbf{d}^{(\tau)} = 0$$

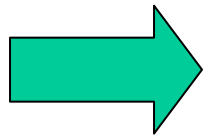
Conjugate gradients 法(3)

$E(\mathbf{w})$ の 2次近似を用いて W 個の conjugate vectors $\{\mathbf{d}_i\}_{i=1}^W$ が求まっているとする (\mathbf{H} , \mathbf{b} は一定とする)

$$E(\mathbf{w}) = E_0 + \mathbf{b}^t \mathbf{w} + \frac{1}{2} \mathbf{w}^t \mathbf{H} \mathbf{w} \quad \mathbf{d}_j^t \mathbf{H} \mathbf{d}_i = 0 \quad j \neq i$$

$\{\mathbf{d}_i\}_{i=1}^W$ は \mathbf{H} が正定値ならば線形独立であり

$$\mathbf{w}^* - \mathbf{w}_1 = \sum_{i=1}^W \alpha_i \mathbf{d}_i \quad \mathbf{w}_j = \mathbf{w}_1 + \sum_{i=1}^{j-1} \alpha_i \mathbf{d}_i$$



$$\mathbf{w}_{j+1} = \mathbf{w}_j + \alpha_j \mathbf{d}_j \quad (1) \quad : \mathbf{w} \text{ の更新式}$$

課題3

次の条件を満たすように定まる conjugate vectors $\{\mathbf{d}_i\}_{i=1}^W$ は the Hessian matrix H が正定値であれば線形独立であることを示しなさい.

$$\mathbf{d}_j^t H \mathbf{d}_i = 0 \quad j \neq i$$

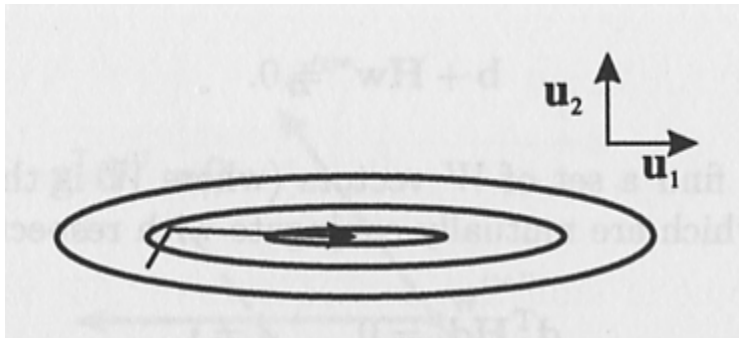
《ヒント》 \mathbf{d}_i for all i が $\{\mathbf{d}_j\}_{j \neq i}$ の線形結合で表せないことが線形独立であることを用いる.

Conjugate gradients 法(4)

$\mathbf{g}(\mathbf{w}) \equiv \nabla E(\mathbf{w}) = \mathbf{b} + \mathbf{H}\mathbf{w}$, $\mathbf{g}(\mathbf{w}^*) = 0$ を用いて

$$\alpha_j = -\frac{\mathbf{d}_j^t \mathbf{g}_j}{\mathbf{d}_j^t \mathbf{H} \mathbf{d}_j} \quad (2) \quad : \mathbf{w} \text{ の更新式の重み係数}$$

$\mathbf{d}_k^t \mathbf{g}_j = 0$ for all $k < j \leq W$ \mathbf{g}_j は $\{\mathbf{d}_k\}_{k=1}^{j-1}$ の全てに直交する



高々 W 回のステップで
最小点 \mathbf{w}^* に到達できる！

Conjugate gradients 法(5)

conjugate vectors $\{\mathbf{d}_i\}_{i=1}^W$ の決定法:

$$\mathbf{d}_1 = -\mathbf{g}_1 (= \mathbf{g}(\mathbf{w}_1)) \quad \mathbf{d}_{j+1} = -\mathbf{g}_{j+1} + \beta_j \mathbf{d}_j \quad (3)$$

$$\beta_j = \frac{\mathbf{g}_{j+1}^t \mathbf{H} \mathbf{d}_j}{\mathbf{d}_j^t \mathbf{H} \mathbf{d}_j} \quad (4) \quad : \mathbf{w}_1 \text{ は random に選択}$$

$$\mathbf{g}_k^t \mathbf{g}_j = 0 \quad \text{for all } k < j \leq W \quad \mathbf{g}_j \text{ は } \{\mathbf{g}_k\}_{k=1}^{j-1} \text{ の全てに直交する}$$

Conjugate gradients 法(6)

$E(w)$ の一般形に対するアルゴリズム:

- (1) 出発点 w_1 を選択する.
- (2) 初期探索方向 $d_1 = -g_1$ を決める.
- (3) 第 j ステップでは, $E(w_j + \alpha d_j)$ を最小化する
 α_{\min} を決定して $w_{j+1} = w_j + \alpha_{\min} d_j$ を定める.
- (4) 最小点に到達したか調べてOKなら終了.
- (5) $g_{j+1} = \nabla E(w_{j+1})$ を求める.
- (6) 新しい探索方向 d_{j+1} を決める.

$$d_{j+1} = -g_{j+1} + \beta_j d_j, \quad \beta_j = \frac{g_{j+1}^t g_{j+1}}{g_j^t g_j}$$

- (7) $j = j+1$ として (3) に戻る.

Newton 法(1)

$E(\mathbf{w})$ の a minimum \mathbf{w}^* の近傍を2次近似すると

$$E(\mathbf{w}) = E(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^t \mathbf{H}(\mathbf{w} - \mathbf{w}^*)$$

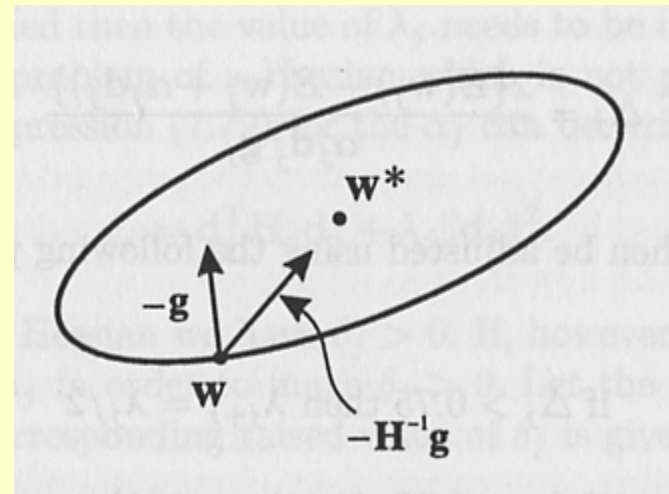
$$\mathbf{g} = \nabla E = \mathbf{H}(\mathbf{w} - \mathbf{w}^*)$$

$$\therefore \mathbf{w}^* = \mathbf{w} - \mathbf{H}^{-1} \mathbf{g}$$

$-\mathbf{H}^{-1} \mathbf{g}$: Newton direction

cf. Gradient descent 法:

$$\mathbf{H}^{-1} \cong \eta \mathbf{I}$$



Newton 法(2)

$E(\mathbf{w})$ の一般形での問題:

- (1) H を評価する計算量 $\sim O(NW^2)$
- (2) H の逆行列の計算量 $\sim O(W^3)$
- (3) H が正定値でないと E 減少の保証なし

$$\left. \frac{\partial}{\partial \lambda} E(\mathbf{w} + \lambda \mathbf{d}) \right|_{\lambda=0} = \mathbf{d}^t \mathbf{g} = -\mathbf{g}^t \mathbf{H} \mathbf{g} < 0$$

H が正定値となるような Newton direction の修正

$$-(\mathbf{H} + \lambda \mathbf{I})^{-1} \mathbf{g} \rightarrow \Delta w_i = - \left(\left| \frac{\partial^2 E}{\partial w_i^2} \right| + \lambda \right)^{-1} \frac{\partial E}{\partial w_i}$$

λ : 正定数

Quasi-Newton 法

H, H^{-1} を直接評価せずに, w, g のみを用いて
 H^{-1} の近似計算を行う $\rightarrow G(\tau)$ の計算

$$w^{(\tau+1)} - w^{(\tau)} = -H^{-1}(g^{(\tau+1)} - g^{(\tau)}) : \text{quasi-Newton 条件}$$

Broyden-Fletcher-Goldfarb-Shanno (BFGS) 公式:

$$G^{(\tau+1)} = G^{(\tau)} + \frac{pp^t}{p^t v} - \frac{(G^{(\tau)} v)v^t G^{(\tau)}}{v^t G^{(\tau)} v} + (v^t G^{(\tau)} v)uu^t$$

$$p = w^{(\tau+1)} - w^{(\tau)}, \quad v = g^{(\tau+1)} - g^{(\tau)}, \quad u = \frac{p}{p^t v} - \frac{G^{(\tau)} v}{v^t G^{(\tau)} v}$$