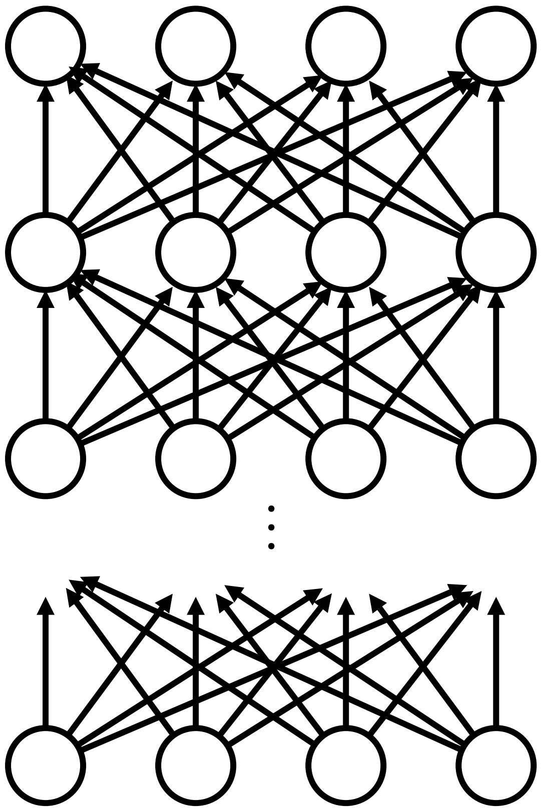
誤差逆伝搬法

MLP のパラメタ推定法:
Error back-propagation



w_{ij}^{1} $\quad o_i^0$: Output layer

w_{ij}^{2} $\quad o_i^1$

: Hidden layers

o_i^2

w_{ij}^{N}

o_i^N : Input layer

$$o_i^{k-1} = f(x_i^k)$$

$$x_i^k = \sum_j o_j^k \cdot w_{ji}^k = \boldsymbol{o}^k \cdot \boldsymbol{w}_i^k$$

$\boldsymbol{o}^N$: 入力

$\boldsymbol{o}^0 = \boldsymbol{g}(\boldsymbol{o}^N; \boldsymbol{w})$: 出力

$\boldsymbol{w}$:モデルパラメタ

$\boldsymbol{d}$: 真値

$$E = \frac{1}{2}\sum_i \left(o_i^0 - d_i\right)^2 = E(\boldsymbol{o}^N, \boldsymbol{d}; \boldsymbol{w})$$: 誤差関数

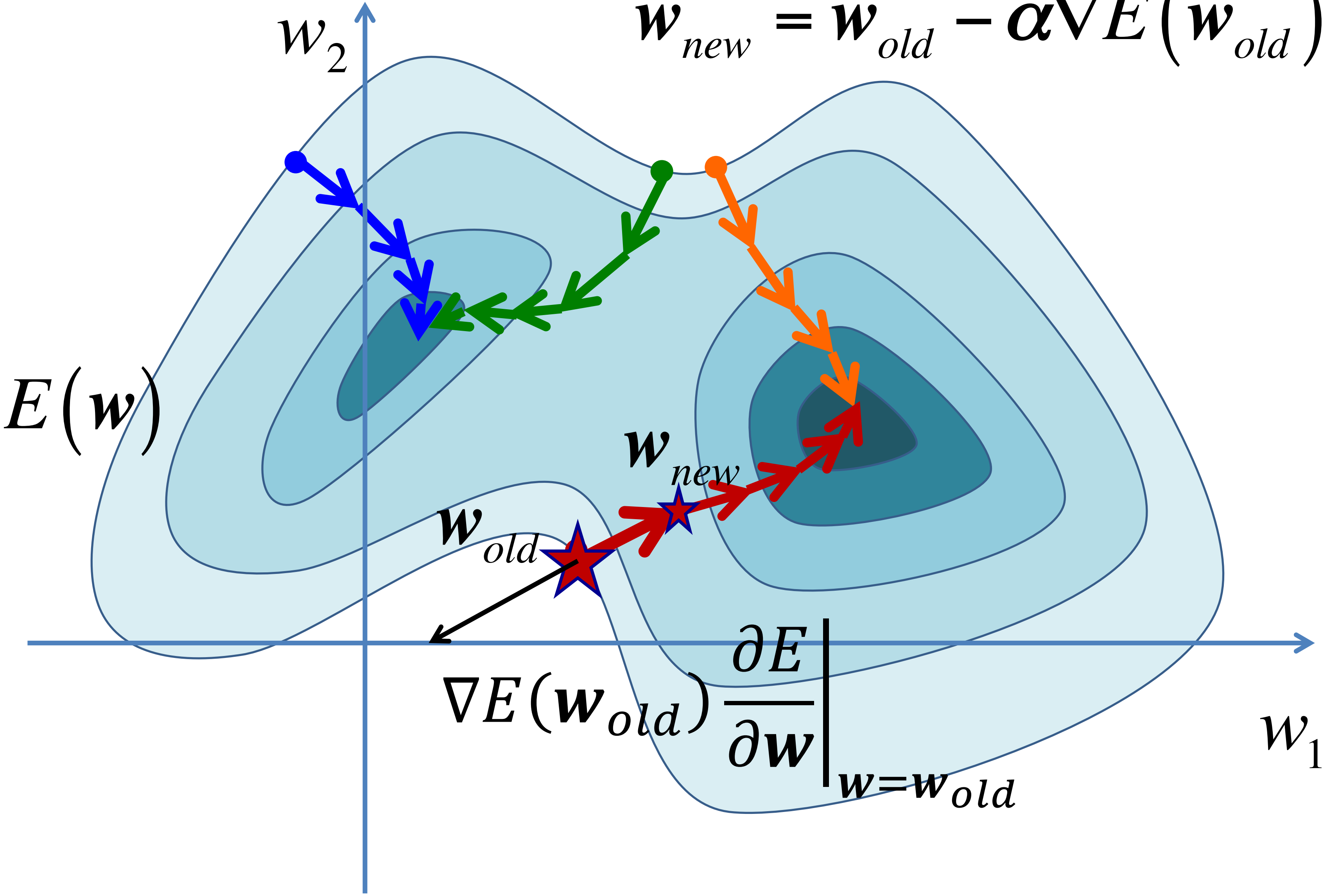
Error back-propagation

= モデルパラメタの関数であるところの誤差関数をモデルパラメタで最小化する

$$w_{ij\_new}^k = w_{ij\_old}^k - \alpha \left.\frac{\partial E}{\partial w_{ij}^k}\right|_{w_{ij}^k = w_{ij\_old}^k}$$

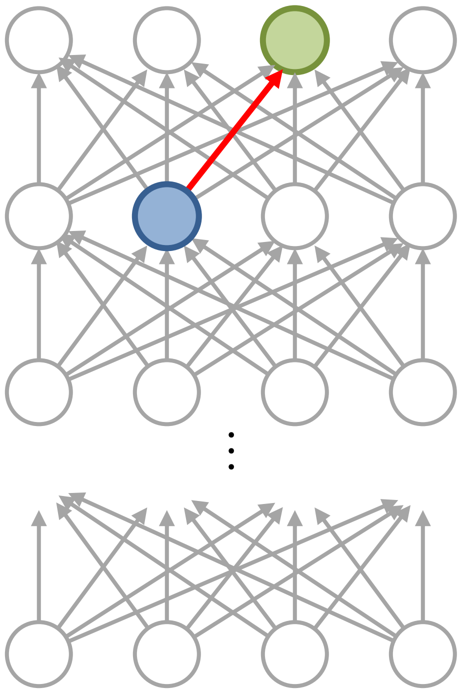
最急降下法

$$\boldsymbol{w}_{new} = \boldsymbol{w}_{old} - \alpha \nabla E\left(\boldsymbol{w}_{old}\right)$$



$$E = \frac{1}{2}\sum_i \left(o_i^0 - d_i\right)^2 = \frac{1}{2}\sum_i \left(f\left(x_i^1\right) - d_i\right)^2$$

$$o_i^0 = f\left(x_i^1\right), \quad x_i^1 = \sum_j o_j^1 w_{ji}^1$$



w_{ji}^1

o_j^1

$$\frac{\partial E}{\partial w_{ji}^1} = \frac{\partial E}{\partial x_i^1}\frac{\partial x_i^1}{\partial w_{ji}^1}$$

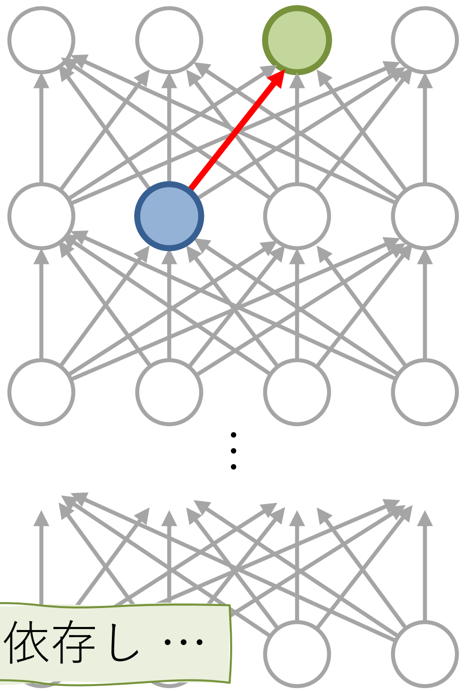
$$\frac{\partial E}{\partial x_i^1} = \left(o_i^0 - d_i\right) f'\left(x_i^1\right) \equiv \delta_i^1$$

$$\frac{\partial x_i^1}{\partial w_{ji}^1} = o_j^1$$

$$\frac{\partial E}{\partial w_{ji}^1} = \delta_i^1 o_j^1, \qquad \hat{w}_{ji}^1 = w_{ji}^1 - \varepsilon\, \delta_i^1 o_j^1$$

$$E = \frac{1}{2}\sum_i \left(o_i^0 - d_i\right)^2 = \frac{1}{2}\sum_i \left(f\left(x_i^1\right) - d_i\right)^2$$

$$o_i^0 = f\left(x_i^1\right),\quad x_i^1 = \sum\nolimits_j o_j^1\, w_{ji}^1$$



$$\frac{\partial E}{\partial w_{ji}^1} = \frac{\partial E}{\partial x_i^1}\frac{\partial x_i^1}{\partial w_{ji}^1}$$

$$\frac{\partial E}{\partial x_i^1} = \left(o_i^0 - d_i\right) f'\left(x_i^1\right) \equiv \delta_i^1$$

Eを減じるために
どれだけ w_{ji}^1 を変えるべきか

は

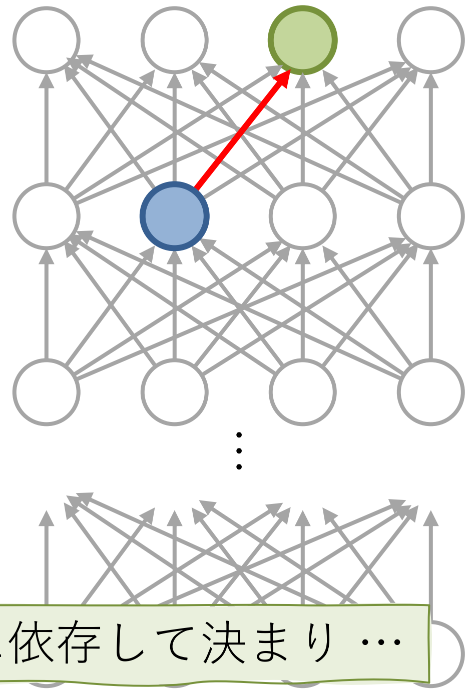
E を減じるために
どれだけ x_i^1 を変えるべきか

に依存し …

$$\frac{\partial E}{\partial w_{ji}^1} = \delta_i^1 o_j^1,\qquad \hat{w}_{ji}^1 = w_{ji}^1 - \varepsilon\, \delta_i^1 o_j^1$$

$$E = \frac{1}{2}\sum_i \left(o_i^0 - d_i\right)^2 = \frac{1}{2}\sum_i \left(f\left(x_i^1\right) - d_i\right)^2$$

$$o_i^0 = f\left(x_i^1\right),\quad x_i^1 = \sum_j o_j^1 w_{ji}^1$$



$$\frac{\partial E}{\partial w_{ji}^1} = \frac{\partial E}{\partial x_i^1}\frac{\partial x_i^1}{\partial w_{ji}^1}$$

$$\frac{\partial E}{\partial x_i^1} = \left(o_i^0 - d_i\right) f'\left(x_i^1\right) \equiv \delta_i^1$$

Eを減じるために
どれだけ w_{ji}^1 を変えるべきか

E を減じるために
どれだけ x_i^1 を変えるべきか

に依存して決まり …

E を減じるために
どれだけ x_i^1 を変えるべきか

$$\frac{\partial E}{\partial w_{ji}^1} = \delta_i^1 o_j^1,$$

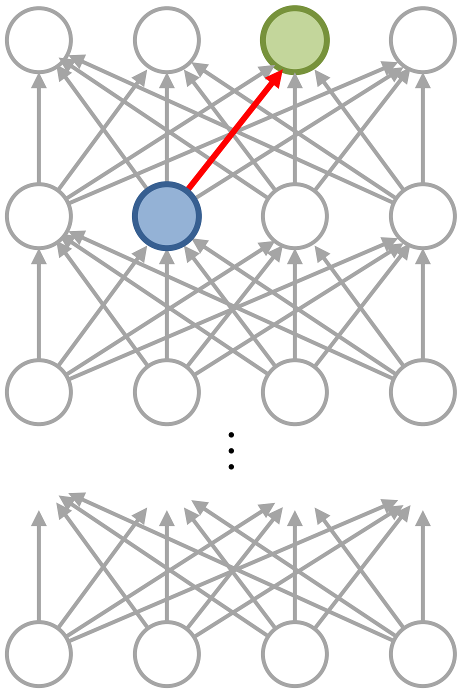
は

Error

に依存して決まる

$$E = \frac{1}{2}\sum_i \left(o_i^0 - d_i\right)^2 = \frac{1}{2}\sum_i \left(f\left(x_i^1\right) - d_i\right)^2$$

$$o_i^0 = f\left(x_i^1\right),\quad x_i^1 = \sum_j o_j^1 w_{ji}^1$$



w_{ji}^1

o_j^1

$$\frac{\partial E}{\partial w_{ji}^1} = \frac{\partial E}{\partial x_i^1}\frac{\partial x_i^1}{\partial w_{ji}^1}$$

$$\frac{\partial E}{\partial x_i^1} = \left(o_i^0 - d_i\right)f'\left(x_i^1\right) \equiv \delta_i^1$$

$$\frac{\partial x_i^1}{\partial w_{ji}^1} = o_j^1$$

E を減じるために
どれだけ x_i^1 を変えるべきか

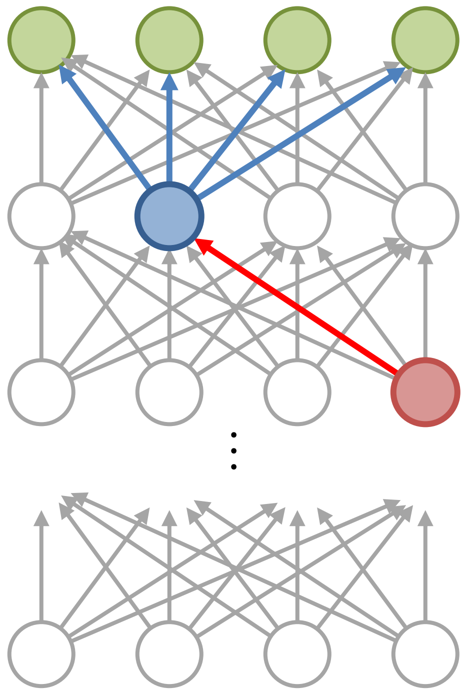
$$\frac{\partial E}{\partial w_{ji}^1} = \delta_i^1 o_j^1,$$

を δ_i^1 とおく

$$\frac{\partial E}{\partial w_{ji}^2} = \frac{\partial E}{\partial x_i^2}\frac{\partial x_i^2}{\partial w_{ji}^2}$$

$$\frac{\partial E}{\partial x_i^2} = \sum_k \frac{\partial E}{\partial x_k^1}\frac{\partial x_k^1}{\partial o_i^1}\frac{\partial o_i^1}{\partial x_i^2}$$

$$= \sum_k \delta_k^1 w_{ik}^1 f'(x_i^2) \equiv \delta_i^2$$



$$\frac{\partial E}{\partial w_{ji}^2} = \delta_i^2 o_j^2$$

$$\widehat{w}_{ji}^2 = w_{ji}^2 - \varepsilon\, \delta_i^2 o_j^2$$

$$\frac{\partial E}{\partial w_{ji}^2} = \frac{\partial E}{\partial x_i^2}\frac{\partial x_i^2}{\partial w_{ji}^2}$$

$$\frac{\partial E}{\partial x_i^2} = \sum_k \frac{\partial E}{\partial x_k^1}\frac{\partial x_k^1}{\partial o_i^1}\frac{\partial o_i^1}{\partial x_i^2}$$

$$= \sum_k \delta_k^1 w_{ik}^1 f'(x_i^2) \equiv \delta_i^2$$

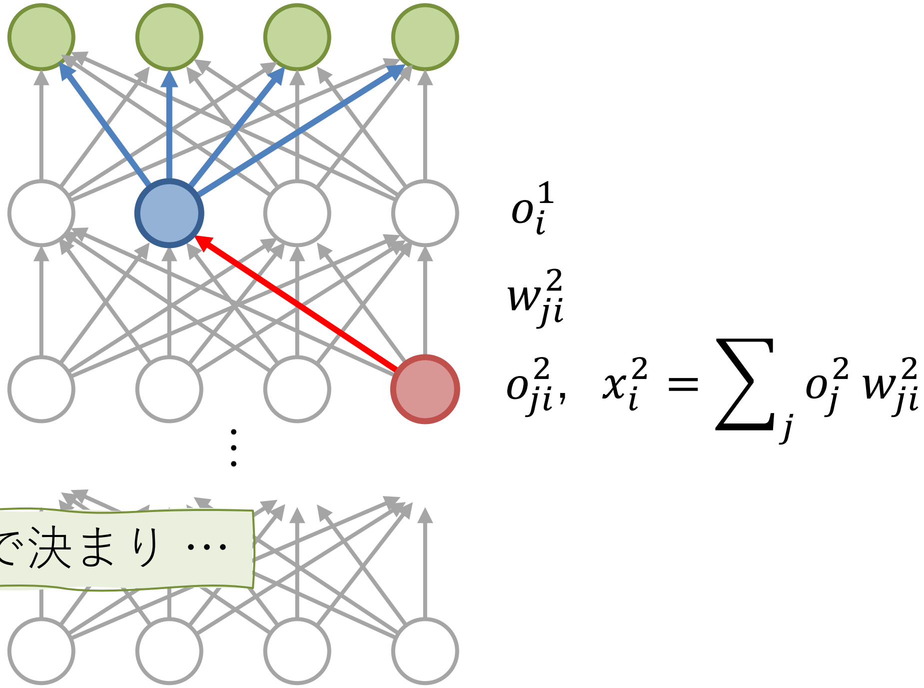
Eを減じるために
どれだけ w_{ji}^2 を変えるべきか

は

Eを減じるために
どれだけ x_i^2 を変えるべきか

で決まり …

$$\frac{\partial E}{\partial w_{ji}^2} = \delta_i^2 o_j^2$$

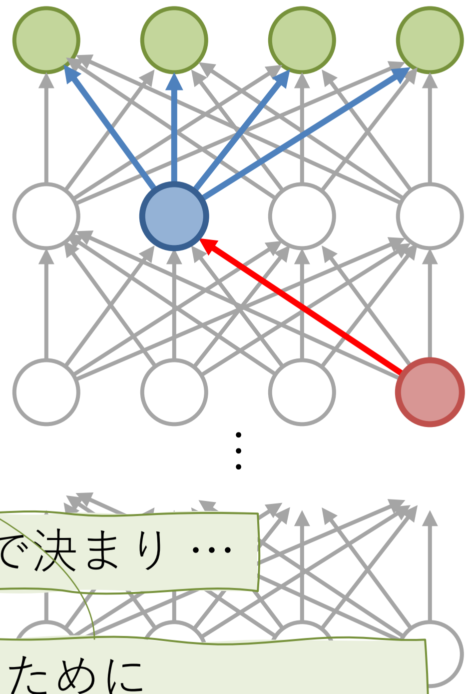


$$\hat{w}_{ji}^2 = w_{ji}^2 - \varepsilon\, \delta_i^2 o_j^2$$

$$\frac{\partial E}{\partial w_{ji}^2} = \frac{\partial E}{\partial x_i^2}\frac{\partial x_i^2}{\partial w_{ji}^2}$$

$$\frac{\partial E}{\partial x_i^2} = \sum_k \frac{\partial E}{\partial x_k^1}\frac{\partial x_k^1}{\partial o_i^1}\frac{\partial o_i^1}{\partial x_i^2}$$

$$= \sum_k \delta_k^1 w_{ik}^1 f'(x_i^2) \equiv \delta_i^2$$



o_i^1

w_{ji}^2

$o_{ji}^2, \quad x_i^2 = \sum_j o_j^2 w_{ji}^2$

Eを減じるために
どれだけ w_{ji}^2 を変えるべきか

は

Eを減じるために
どれだけ x_i^2 を変えるべきか

で決まり …

$$\frac{\partial E}{\partial w_{ji}^2} = \delta_i^2 o_j^2$$

Eを減じるために
どれだけ x_i^2 を変えるべきか

は

Eを減じるために
どれだけ x_i^1 を変えるべきか

で決まる

Eを減じるために
どれだけ x_i^1 を変えるべきか

は

Error

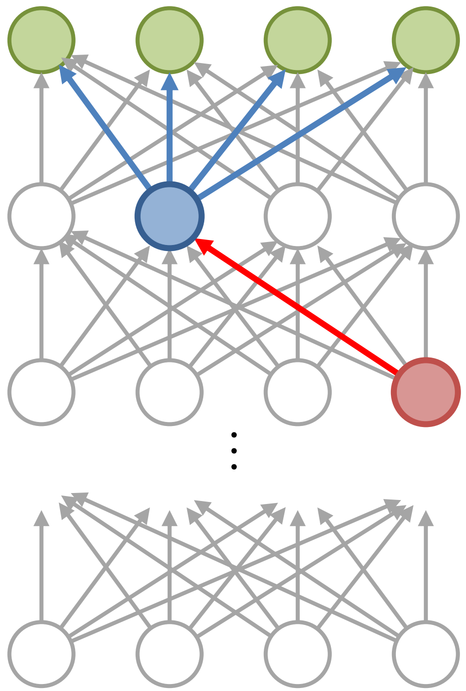
で決まる

$$\hat{w}_{ji}^2 = w_{ji}^2 - \varepsilon\, \delta_i^2 o_j^2$$

$$\frac{\partial E}{\partial w_{ji}^2} = \frac{\partial E}{\partial x_i^2}\frac{\partial x_i^2}{\partial w_{ji}^2}$$

$$\frac{\partial E}{\partial x_i^2} = \sum_k \frac{\partial E}{\partial x_k^1}\frac{\partial x_k^1}{\partial o_i^1}\frac{\partial o_i^1}{\partial x_i^2}$$

$$= \sum_k \delta_k^1 w_{ik}^1 f'(x_i^2) \equiv \delta_i^2$$



$$\frac{\partial E}{\partial w_{ji}^2} = \delta_i^2 o_j^2$$

Eを減じるために
どれだけ x_i^2 を変えるべきか は δ_i^1 で決まる

$$\hat{w}_{ji}^2 = w_{ji}^2 - \varepsilon\, \delta_i^2 o_j^2$$

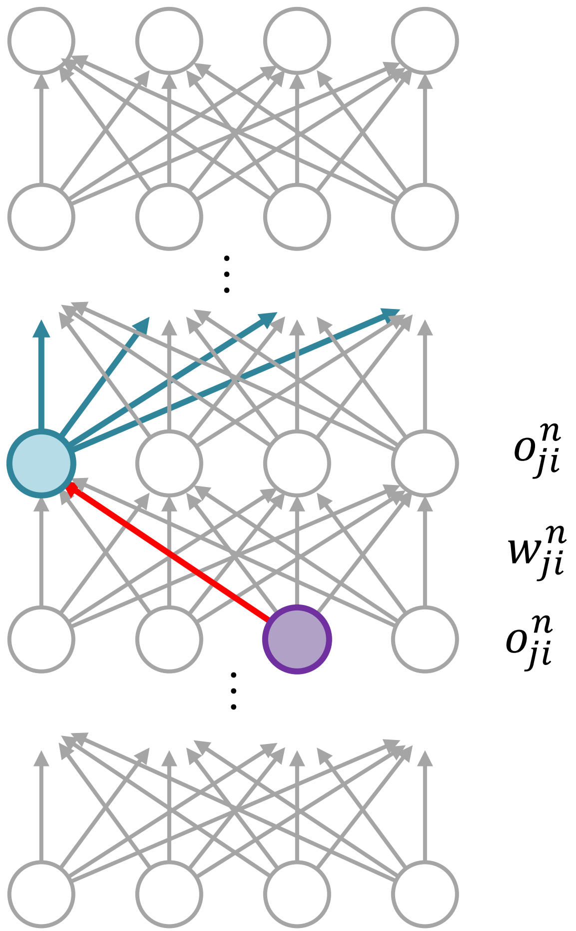
$$\frac{\partial E}{\partial w_{ji}^n} = \frac{\partial E}{\partial x_i^n}\frac{\partial x_i^n}{\partial w_{ji}^n}$$

$$\frac{\partial E}{\partial x_i^n} = \sum_k \frac{\partial E}{\partial x_k^{n-1}}\frac{\partial x_k^{n-1}}{\partial o_i^{n-1}}\frac{\partial o_i^{n-1}}{\partial x_i^n}$$

$$= \sum_k \delta_k^{n-1} w_{ik}^{n-1} f'(x_i^n) \equiv \delta_i^n$$

$$\frac{\partial E}{\partial w_{ji}^n} = \delta_i^n o_j^n$$

$$\widehat{w}_{ji}^n = w_{ji}^n - \varepsilon\, \delta_i^n o_j^n$$



$$\frac{\partial E}{\partial w_{ji}^n} = \frac{\partial E}{\partial x_i^n}\frac{\partial x_i^n}{\partial w_{ji}^n}$$

$$\frac{\partial E}{\partial x_i^n} = \sum_k \frac{\partial E}{\partial x_k^{n-1}}\frac{\partial x_k^{n-1}}{\partial o_i^{n-1}}\frac{\partial o_i^{n-1}}{\partial x_i^n}$$

Eを減じるために
どれだけ w_{ji}^n を変えるべきか

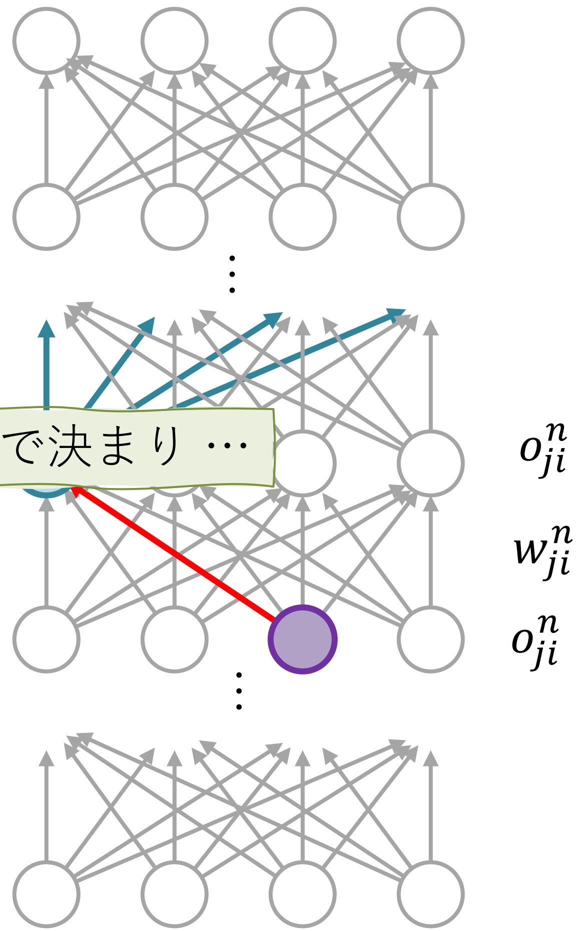
は

Eを減じるために
どれだけ x_i^n を変えるべきか

で決まり …

$$\frac{\partial E}{\partial w_{ji}^n} = \delta_i^n o_j^n$$

$$\widehat{w}_{ji}^n = w_{ji}^n - \varepsilon\, \delta_i^n o_j^n$$



$$\frac{\partial E}{\partial w_{ji}^n} = \frac{\partial E}{\partial x_i^n}\frac{\partial x_i^n}{\partial w_{ji}^n}$$

$$\frac{\partial E}{\partial x_i^n} = \sum_k \frac{\partial E}{\partial x_k^{n-1}}\frac{\partial x_k^{n-1}}{\partial o_i^{n-1}}\frac{\partial o_i^{n-1}}{\partial x_i^n}$$

Eを減じるために
どれだけ w_{ji}^n を変えるべきか

は

Eを減じるために
どれだけ x_i^n を変えるべきか

で決まり …

Eを減じるために
どれだけ x_i^n を変えるべきか

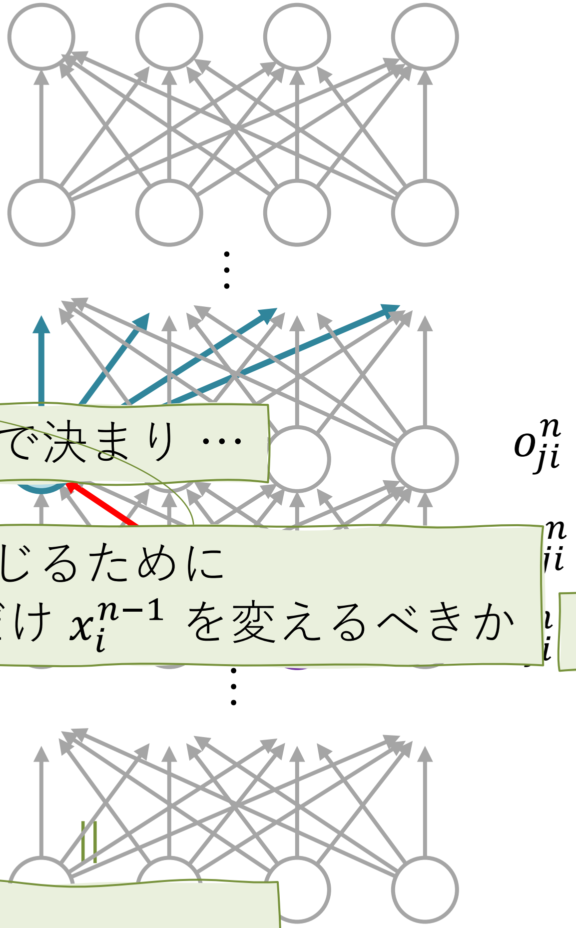
は

Eを減じるために
どれだけ x_i^{n-1} を変えるべきか

で決まり …

$$\frac{\partial E}{\partial w_{ji}^n} = \delta_i^n o_j^n$$

$$\hat{w}_{ji}^n = w_{ji}^n - \varepsilon\, \delta_i^n o_j^n$$



Eを減じるために
どれだけ x_i^1 を変えるべきか

は

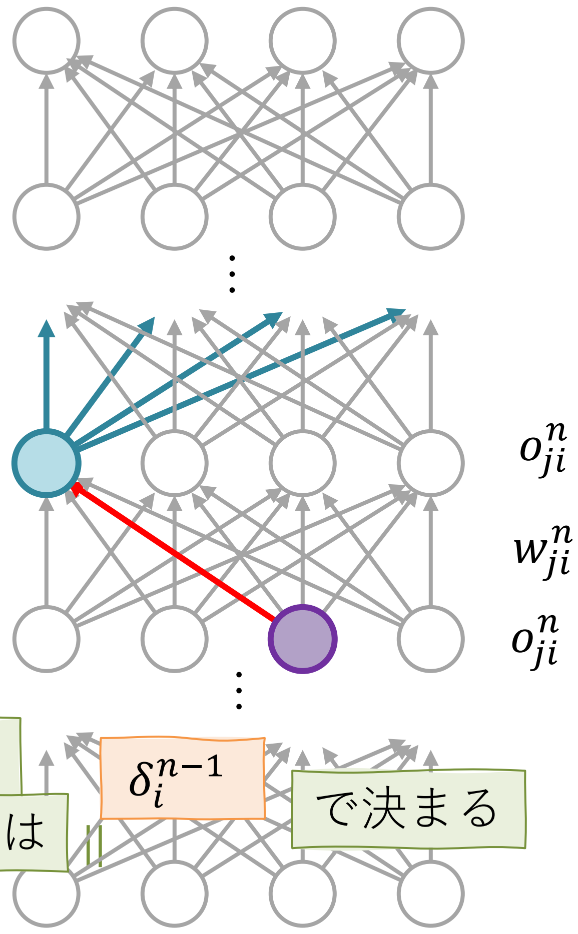
Error

で決まる

$$\frac{\partial E}{\partial w_{ji}^{n}} = \frac{\partial E}{\partial x_{i}^{n}} \frac{\partial x_{i}^{n}}{\partial w_{ji}^{n}}$$

$$\frac{\partial E}{\partial x_{i}^{n}} = \sum_{k} \frac{\partial E}{\partial x_{k}^{n-1}} \frac{\partial x_{k}^{n-1}}{\partial o_{i}^{n-1}} \frac{\partial o_{i}^{n-1}}{\partial x_{i}^{n}}$$

$$= \sum_{k} \delta_{k}^{n-1} w_{ik}^{n-1} f'(x_{i}^{n}) \equiv \delta_{i}^{n}$$



$$\frac{\partial E}{\partial w_{ji}^{n}} = \delta_{i}^{n} o_{j}^{n}$$

$$\hat{w}_{ji}^{n} = w_{ji}^{n} - \varepsilon\, \delta_{i}^{n} o_{j}^{n}$$

Eを減じるために
どれだけ x_i^n を変えるべきか は δ_i^{n-1} で決まる

MLPのパラメタ推定が抱える問題

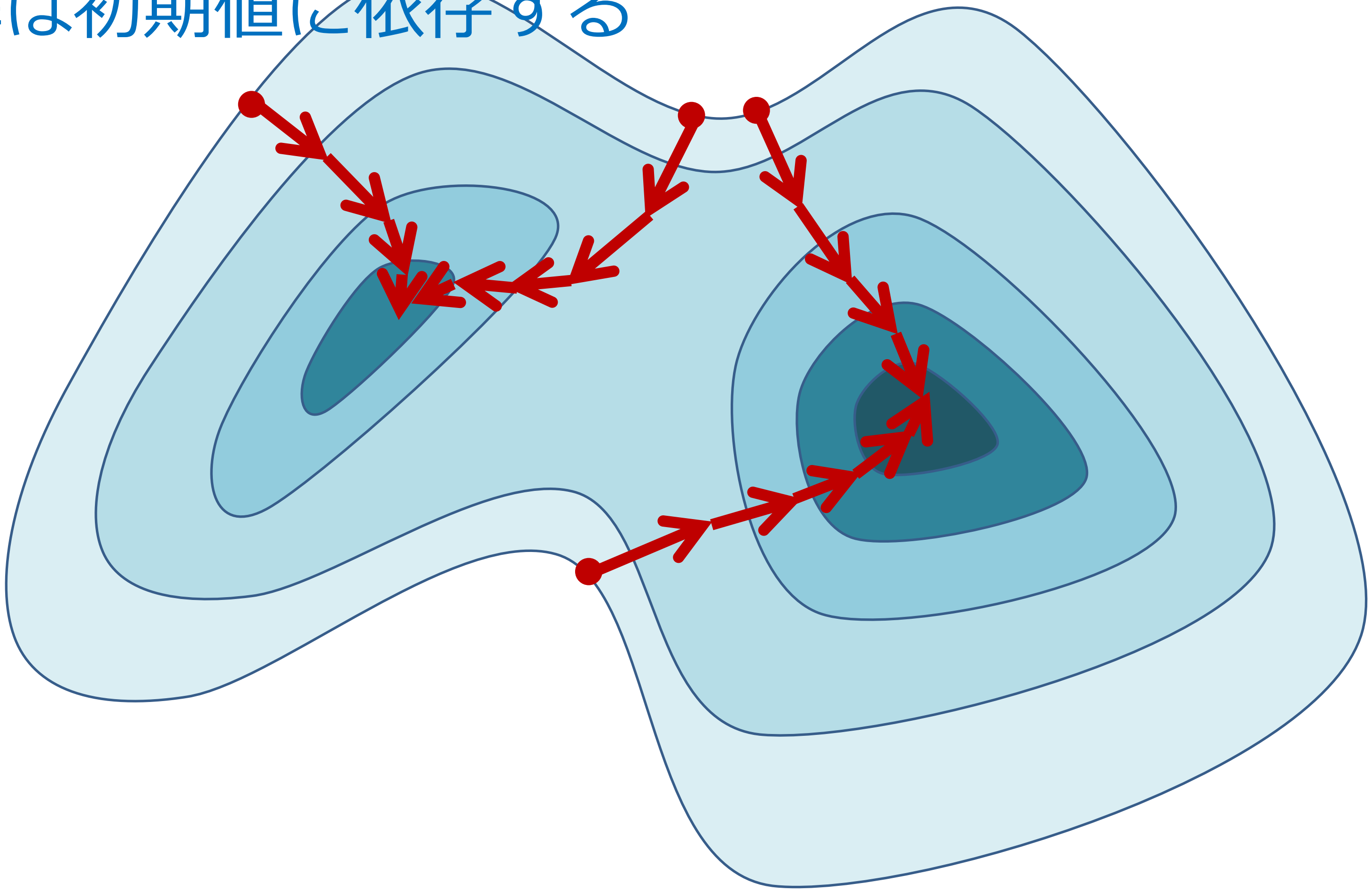
パラメタ推定における問題

- 勾配法が抱える一般的問題
  - 解の初期値依存性
    - 局所最適解にトラップされる
    - 勾配消失問題
  - 不定解を生じやすい

- 勾配法が一般的に抱える問題は，DNNの構造が複雑なだけに顕在化しやすい。

勾配法と初期値問題

勾配法においては,
得られる解は初期値に依存する

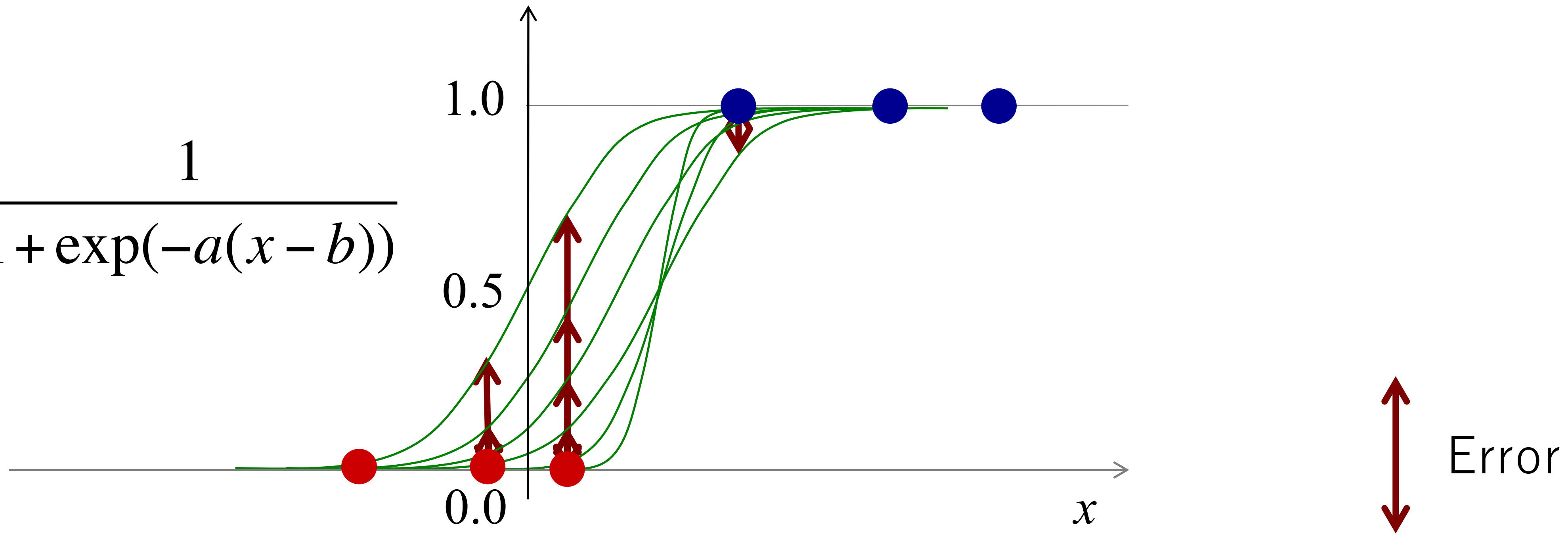


勾配消失問題

1-dimensional view

初期解が最適解に近ければ，最適解をみつけて収束する。

$$f(x)=\frac{1}{1+\exp(-a(x-b))}$$

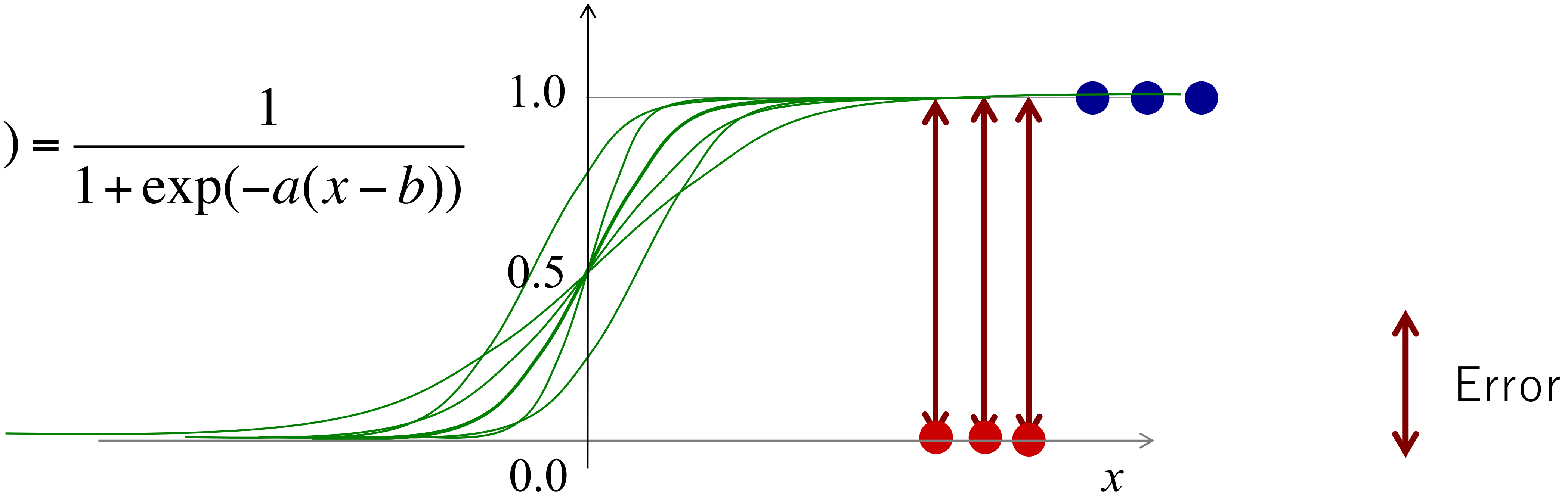


勾配消失問題

1-dimensional view

初期解が最適解から離れているとき，
解を見つけることなくパラメタ更新が止まる。

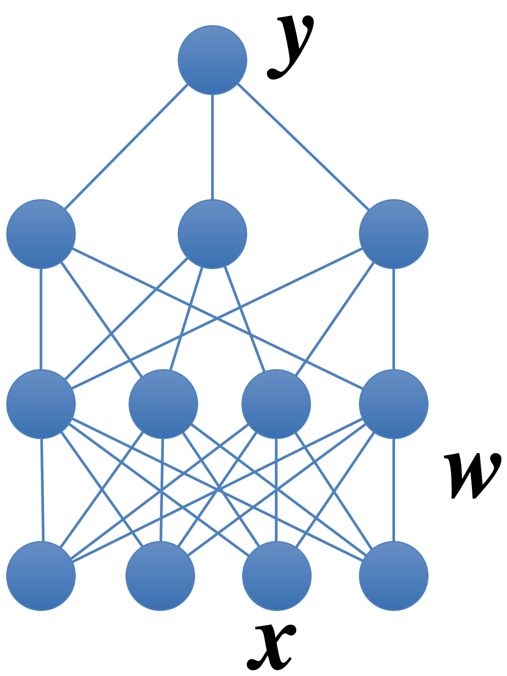
$$f(x) = \frac{1}{1 + \exp(-a(x-b))}$$

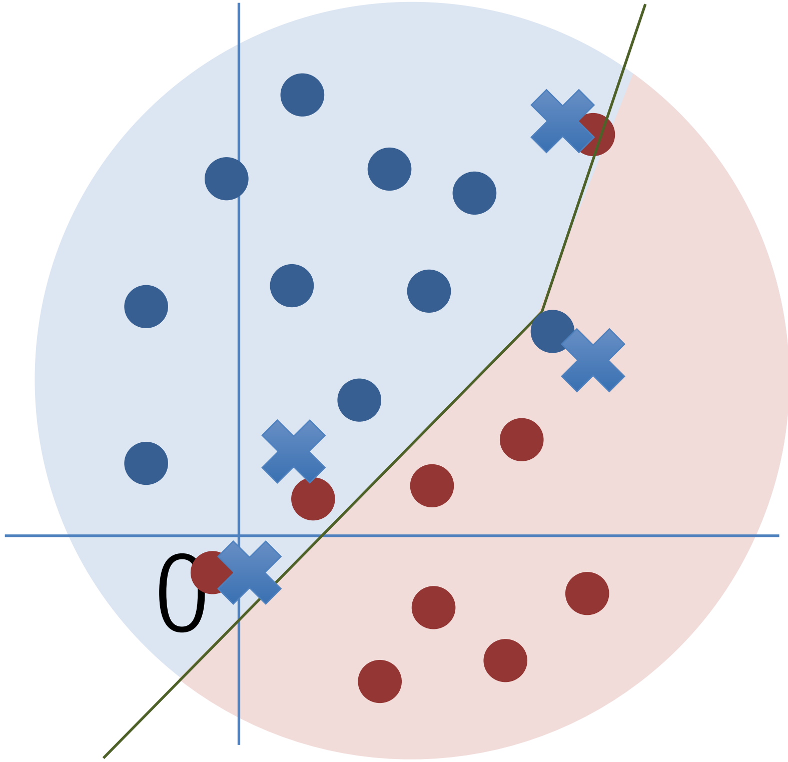


勾配消失問題

2-dimensional view

初期解が最適解に近ければ，

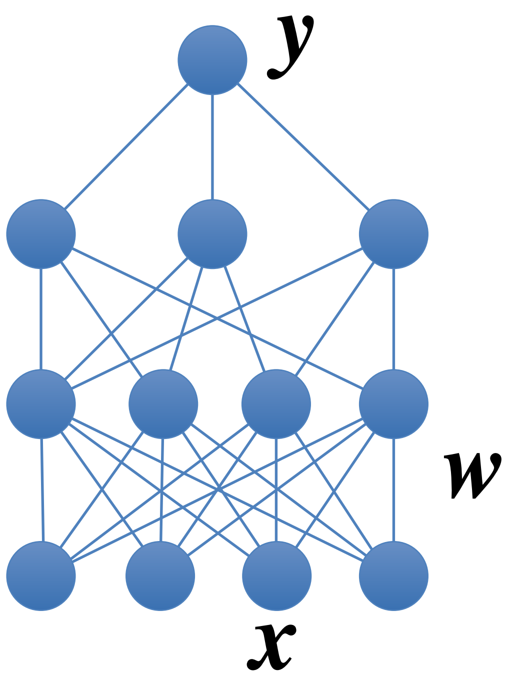


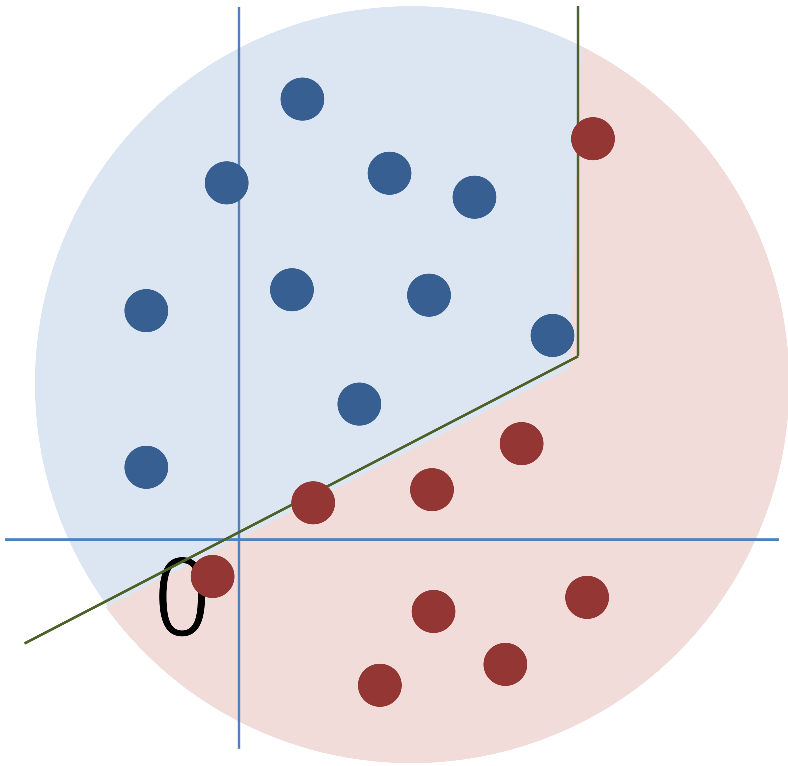


勾配消失問題

2-dimensional view

初期解が最適解に近ければ，最適解をみつけて収束する。

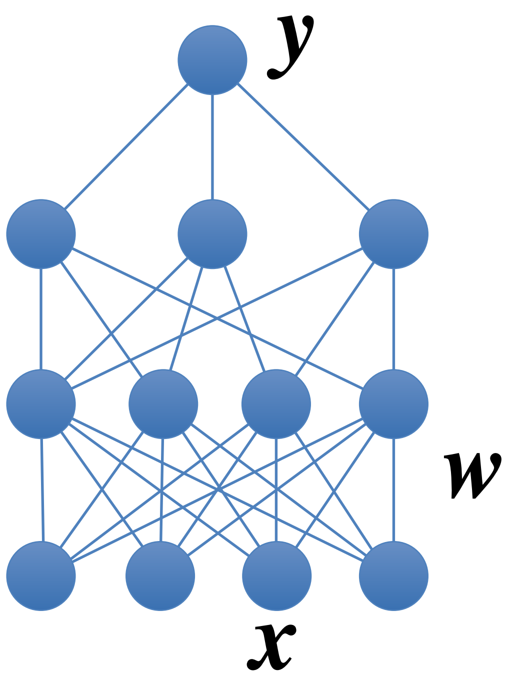


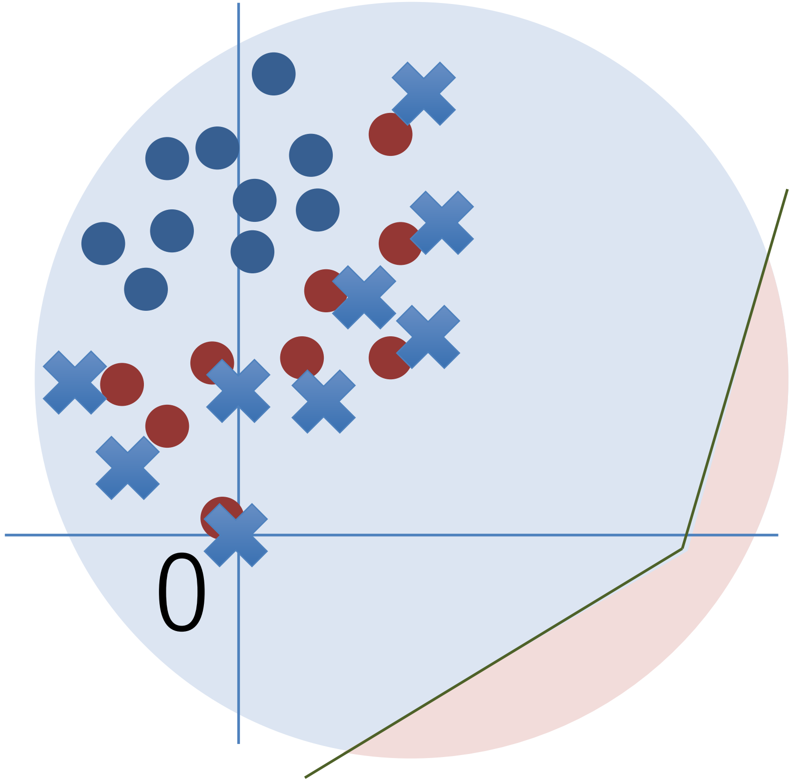


勾配消失問題

2-dimensional view

初期解が最適解から離れているとき,

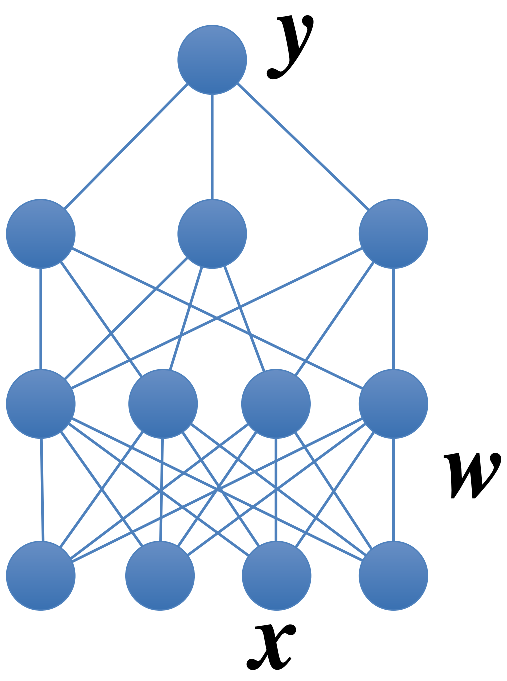


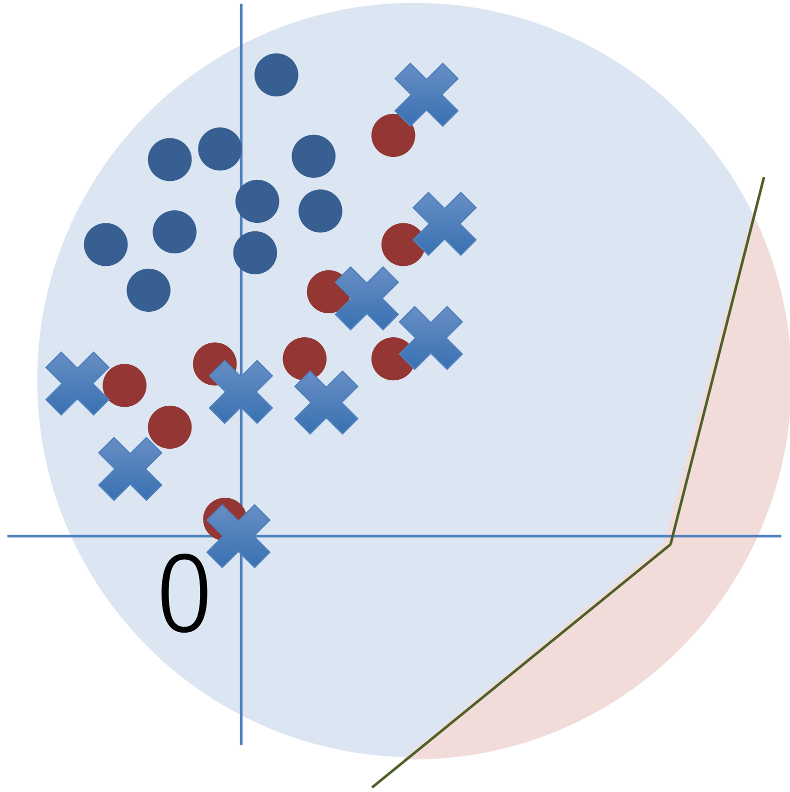


勾配消失問題

2-dimensional view

初期解が最適解から離れているとき，
解を見つけることなくパラメタ更新が止まる。

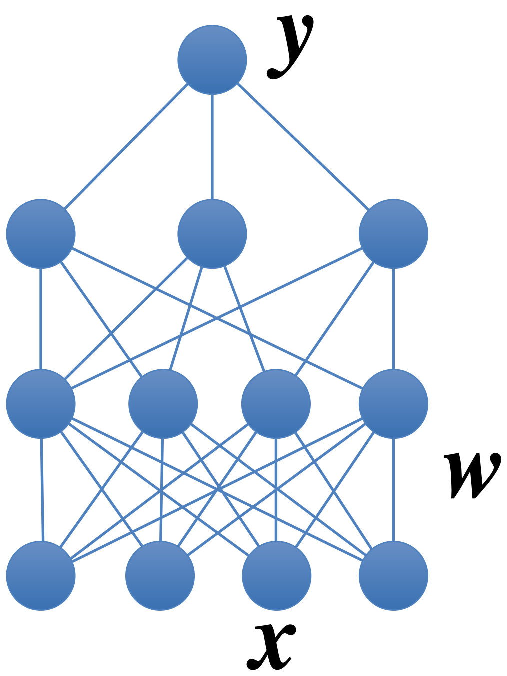


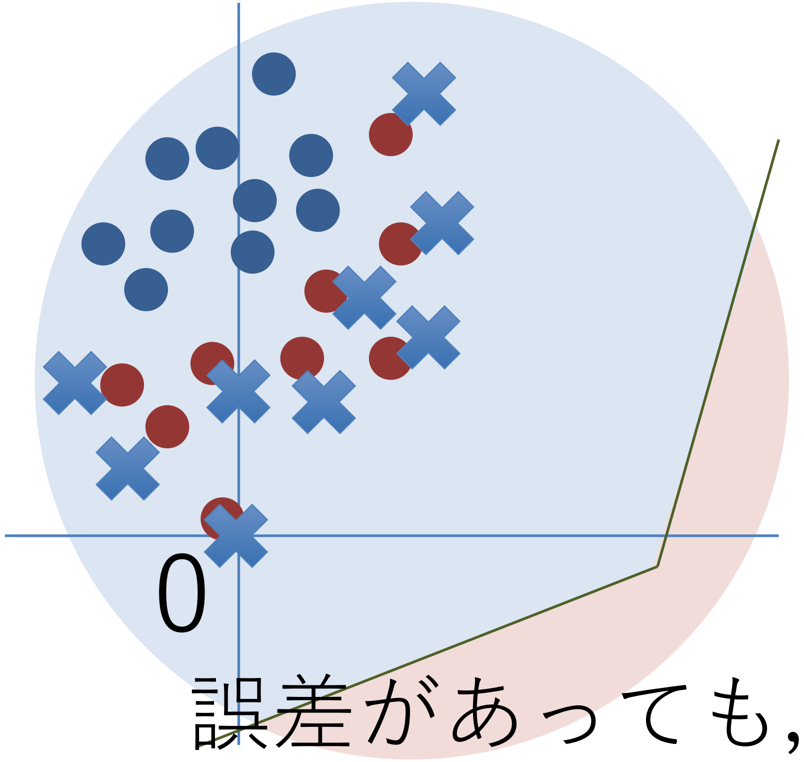


勾配消失問題

2-dimensional view

初期解が最適解から離れているとき，
解を見つけることなくパラメタ更新が止まる。





$$\left.\frac{\partial E}{\partial \boldsymbol{w}}\right|_{old} = 0$$

$$\boldsymbol{w}_{new} = \boldsymbol{w}_{old}$$

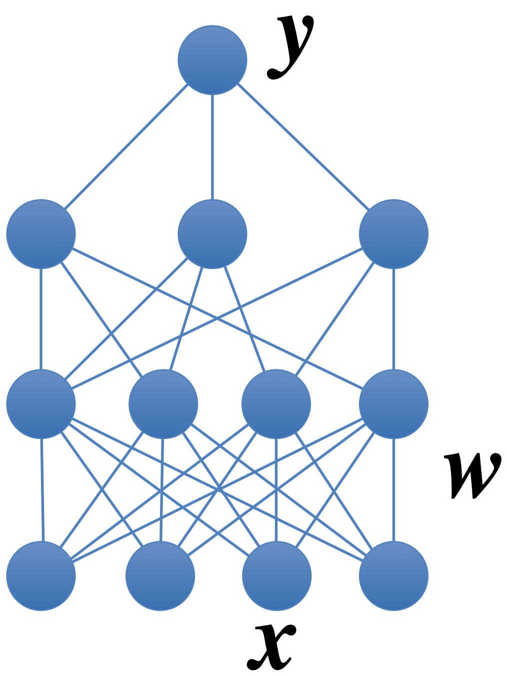
誤差があっても，傾きがゼロ

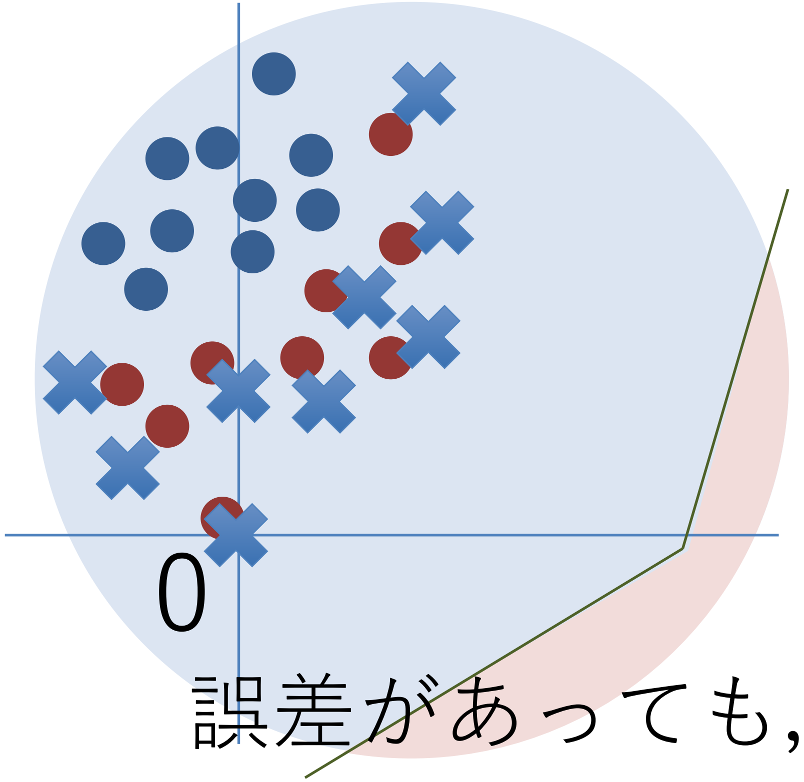
→パラメタは更新されない

勾配消失問題

2-dimensional view

初期解が最適解から離れているとき，
解を見つけることなくパラメタ更新が止まる。





$$\left.\frac{\partial E}{\partial \boldsymbol{w}}\right|_{old} = 0$$

$$\boldsymbol{w}_{new} = \boldsymbol{w}_{old}$$

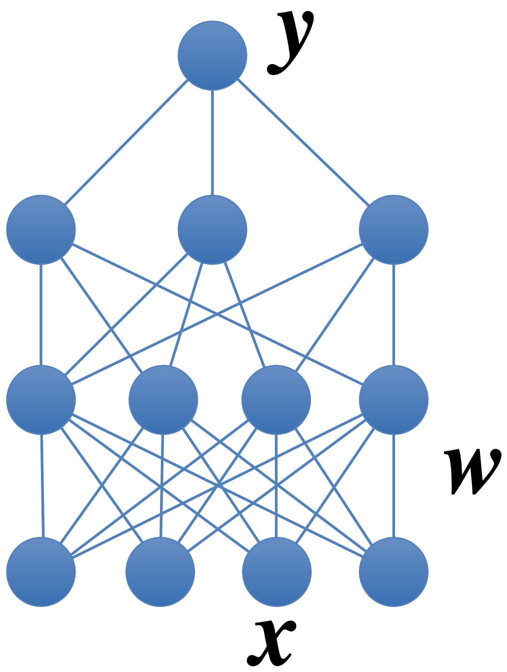
誤差があっても，傾きがゼロ

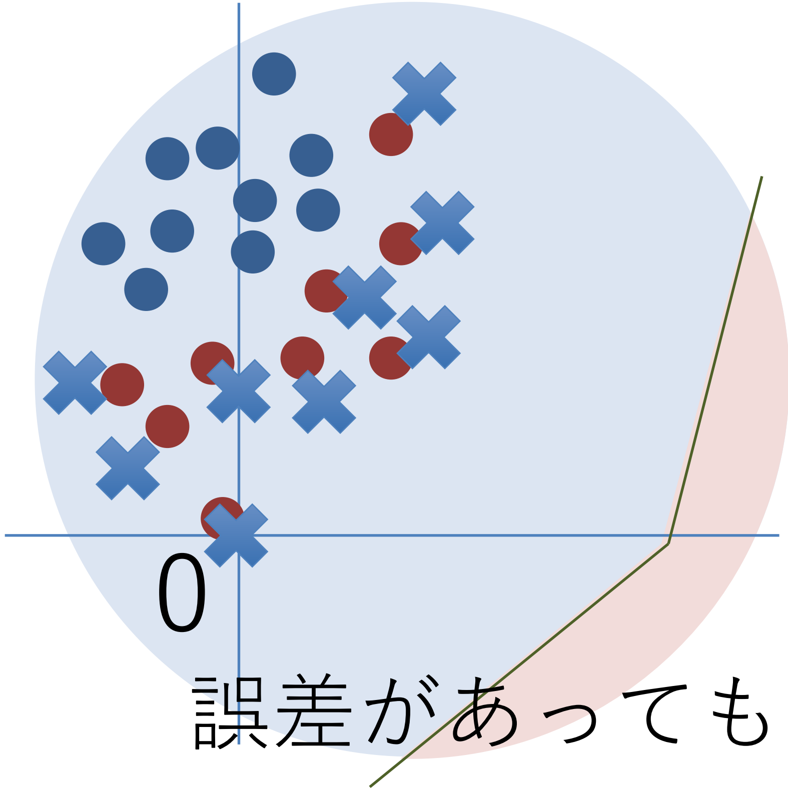
→パラメタは更新されない

勾配消失問題

2-dimensional view

初期解が最適解から離れているとき，
解を見つけることなくパラメタ更新が止まる。





$$\left.\frac{\partial E}{\partial \boldsymbol{w}}\right|_{old} = 0$$

$$\boldsymbol{w}_{new} = \boldsymbol{w}_{old}$$

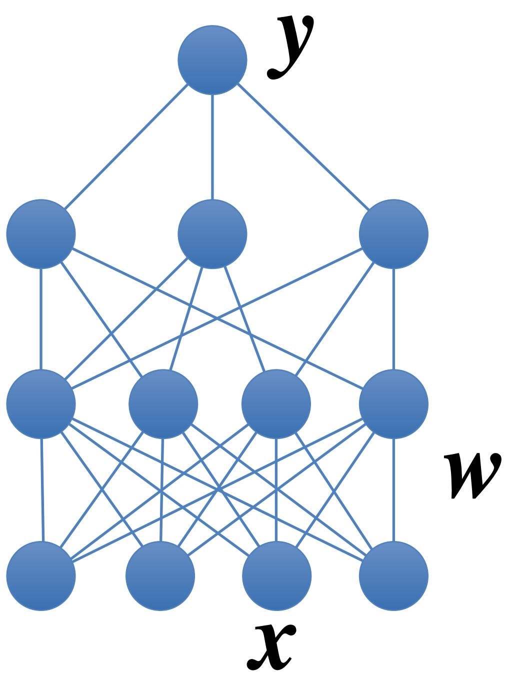
誤差があっても，傾きがゼロ

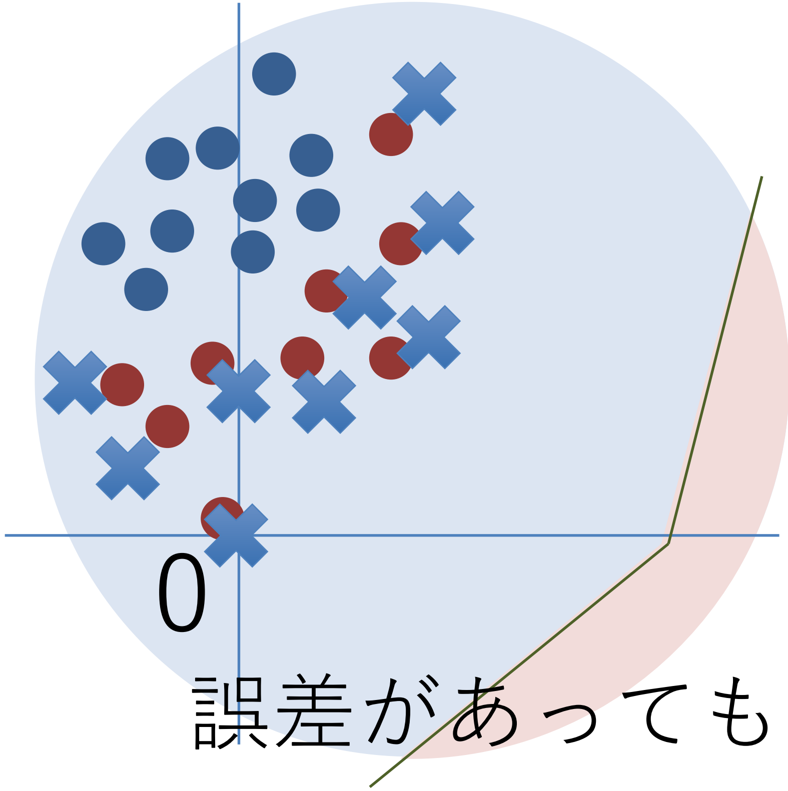
→パラメタは更新されない

勾配消失問題

2-dimensional view

初期解が最適解から離れているとき，
解を見つけることなくパラメタ更新が止まる。





$$\left.\frac{\partial E}{\partial \boldsymbol{w}}\right|_{old} = 0$$

$$\boldsymbol{w}_{new} = \boldsymbol{w}_{old}$$

誤差があっても，傾きがゼロ

Vanishing gradient problem

→パラメタは更新されない

まとめ

- MLPのパラメタ推定には，誤差逆伝播法が用いられる。
- 誤差逆伝播法は，勾配法を採用する。
- 誤差逆伝播法は，初期値問題，勾配消失問題などの問題を持つ。