$(x_i, y_i)_{i=1}^n$

$x_i \in \mathbb{R}^m$

$x_{i,j}$ j^{th} feature of i^{th} point.

$$\hat{y}_i = f_w(x_i) = w^T x_i = \sum_{j=1}^{m} w_j x_{i,j}$$

$$r_i = y_i - \hat{y}_i$$

$$l(w) = \sum_{i=1}^{n} (r_i)^2 = \sum_{i=1}^{n} (y_i - f_w(x_i))^2$$

$$= \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j x_{i,j}\right)^2$$

$w^* \in \underset{w}{\operatorname{argmin}}\ l(w)$

Hill Climbing.

sequence of model parameters $w_i \in \mathbb{R}^m$

w_0

w_1

w_2

$\vdots$

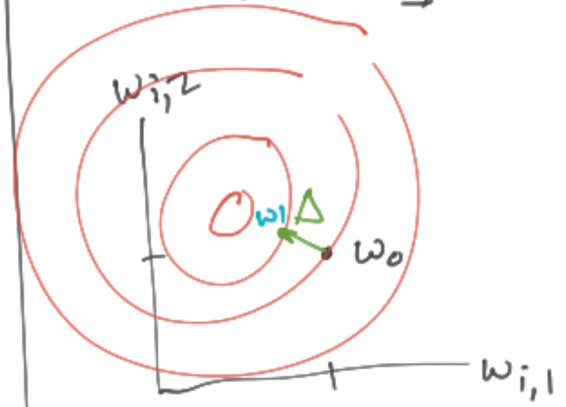
$w_{i,j}$ → which weight of w_i.

↓ which term in sequence

Lecture 5

- Gradient descent
- Basis functions
- Feature normalization.

"Directions"

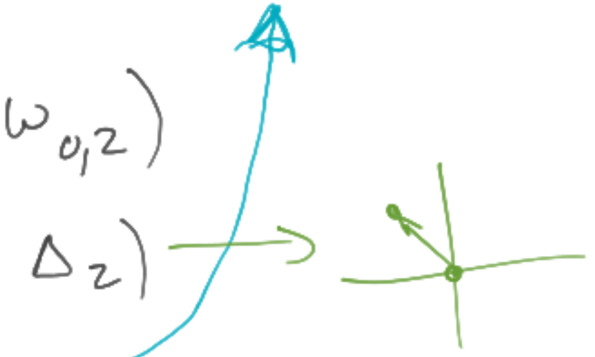


$$w_1 = (w_{0,1} + \Delta_1, w_{0,2} + \Delta_2)$$

$$w_0 = (w_{0,1}, w_{0,2})$$

$$\Delta = (\Delta_1, \Delta_2)$$

$$w_1 = w_0 + \Delta$$



Dir. Steepest Ascent of l at w_i is gradient:

$$\nabla l(w_i) = \left(\frac{\partial l(w_i)}{\partial w_{i,1}} , \frac{\partial l(w_i)}{\partial w_{i,2}} , \dots , \frac{\partial l(w_i)}{\partial w_{i,m}} \right)$$

or random.
often zero.

$-\nabla l(w_i)$ is the dir. of steepest descent.

Gradient Descent

Hyperparams:
w_0 : initial point
α : stepsize.
stopping criteria.

$x_i \in \mathbb{R}^m$

$f_{w_i}(x) = \sum_{j=1}^{m} w_{i,j} x_j$

Input: loss function $l : \mathbb{R}^m \to \mathbb{R}$
dimension m of w

Output: An approximation of an element of $\arg\min_w l(w)$

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \hat{y}_i \to f_w(x_i)$$

$i \leftarrow 0$

while stopping criteria not satisfied do

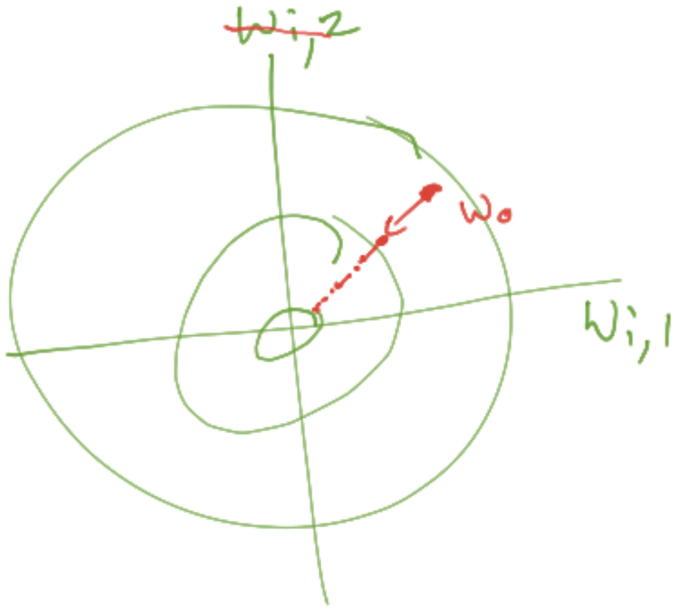
iter.

  for $j = 1$ to m do

    $w_{i+1,j} = w_{i,j} - \alpha \frac{\partial \ell(w_i)}{\partial w_{i,j}}$

  $i = i + 1$

return w_i

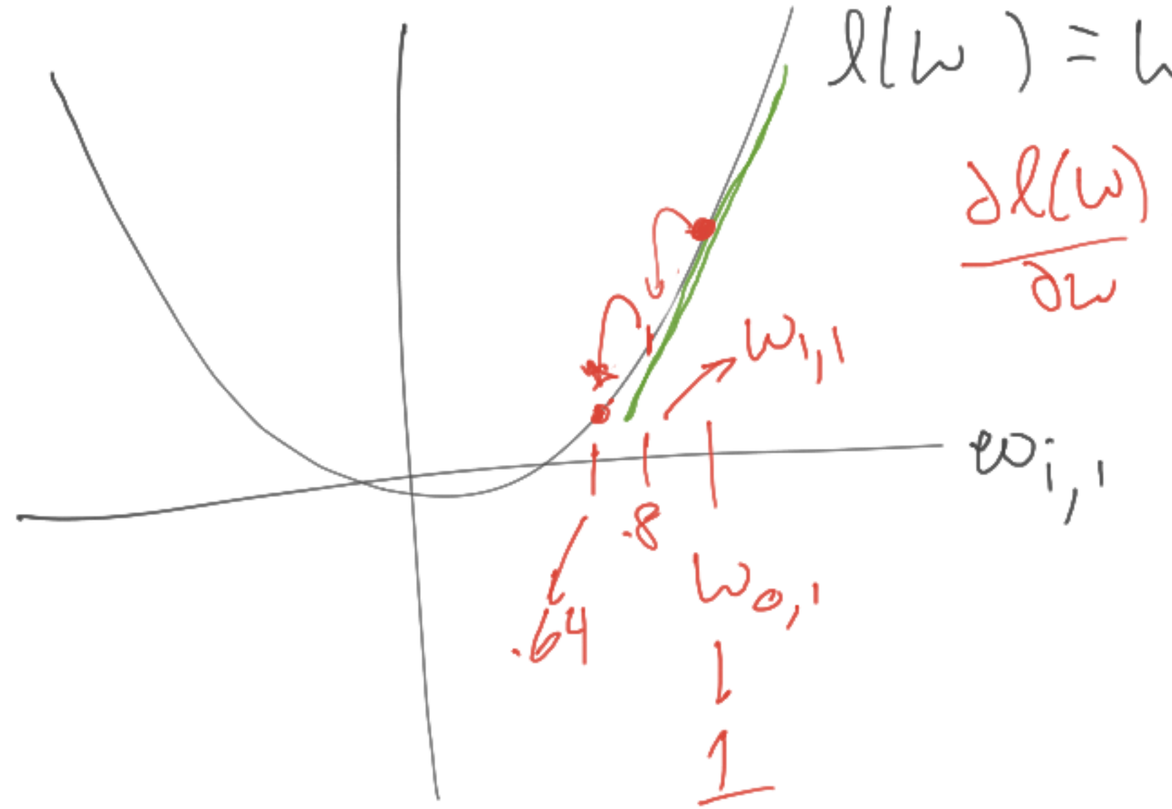


→ After fixed time.

→ When $\ell(w_i)$ sufficiently small

→ When $D\ell(w_i)$ is small.

→ When $\ell(w_i)$ not much less than $\ell(w_{i-k})$

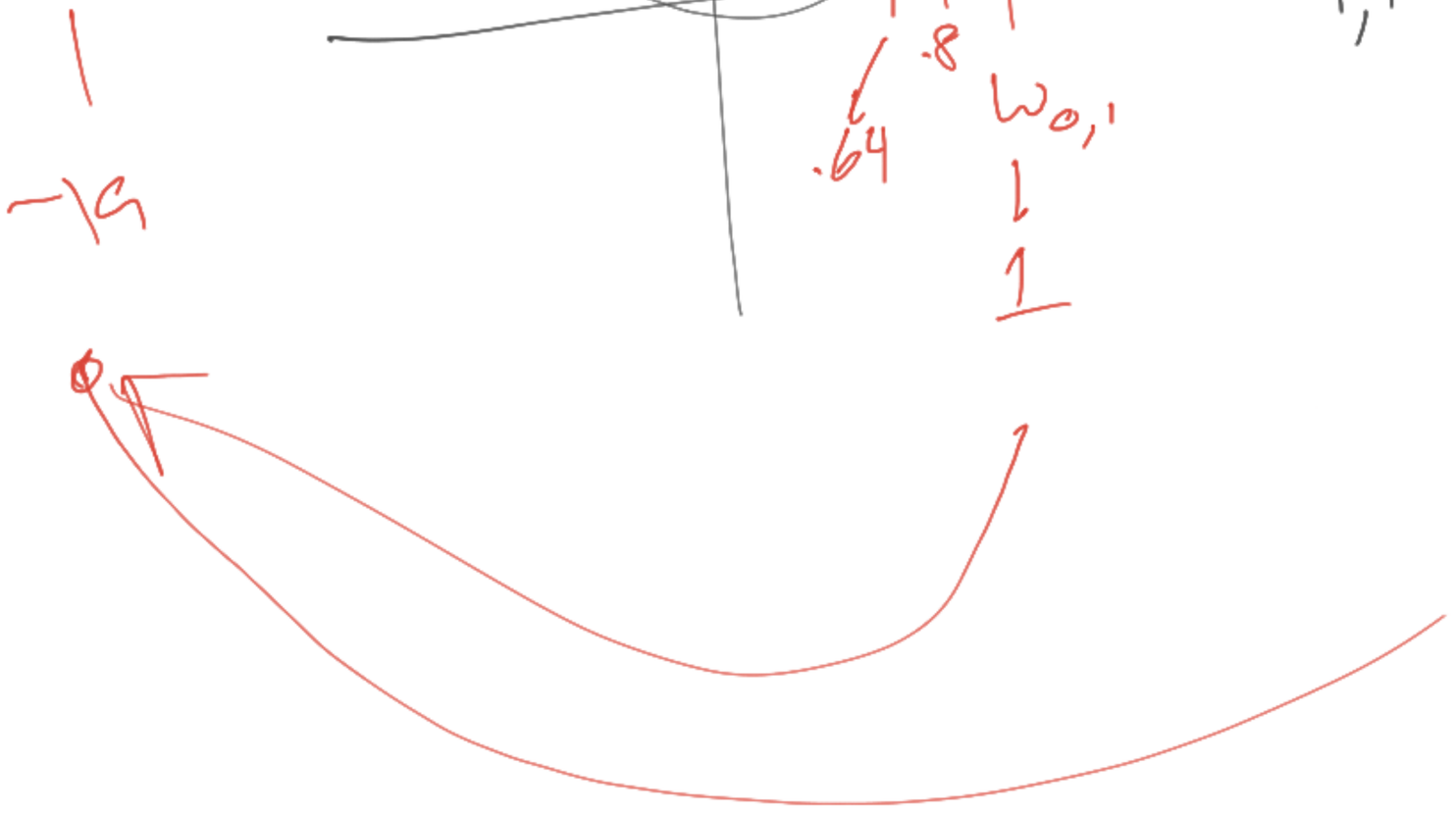


$l(w) = w^2$ $\quad w \in \mathbb{R}$ $\quad .8 - .1(1.6)$

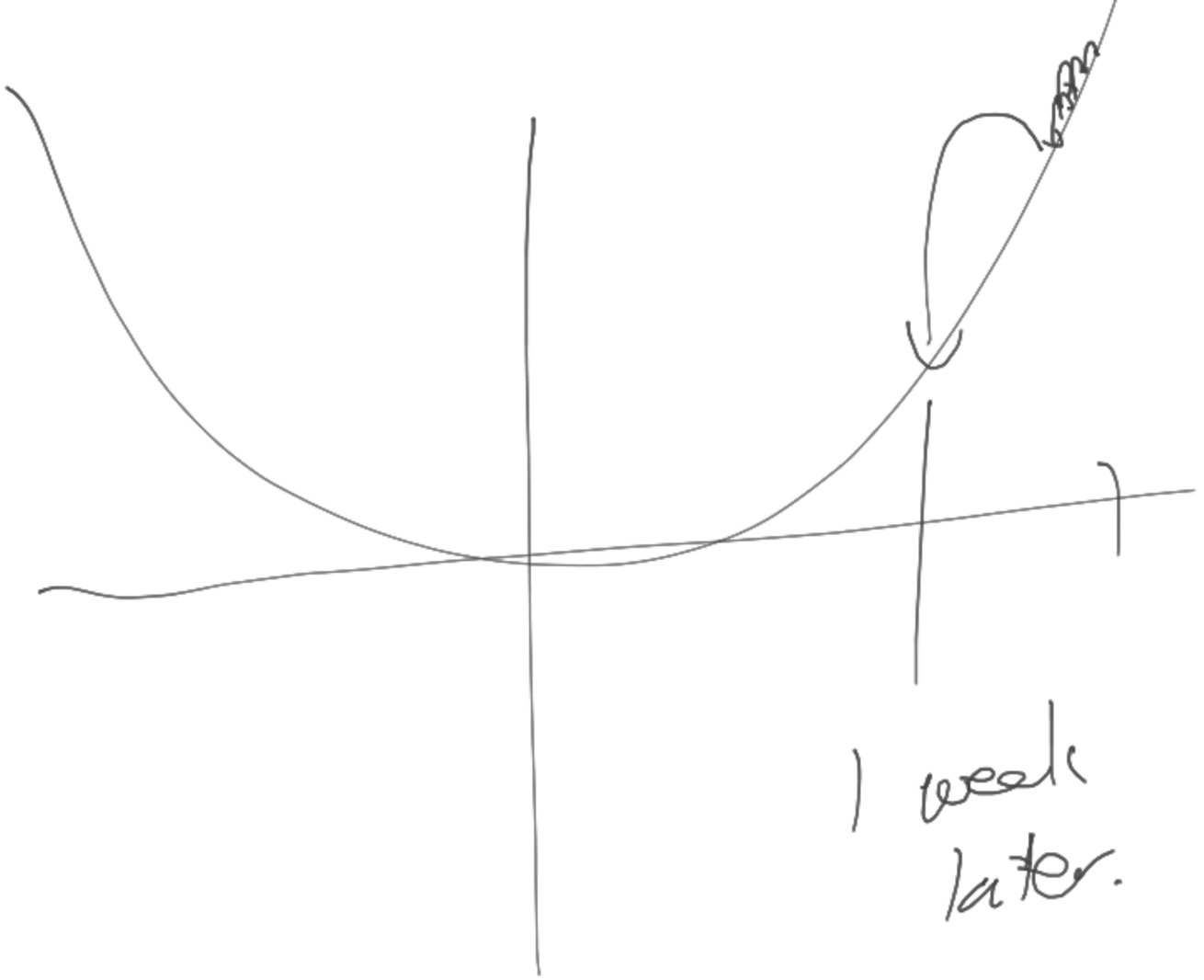
$\frac{\partial l(w)}{\partial w} = 2w$ $\quad .8 - .1(2 \cdot .8)$

$w_{i+1,1} = w_{i,1} - \alpha(2w_{i,1})$

$\alpha = 0.1$

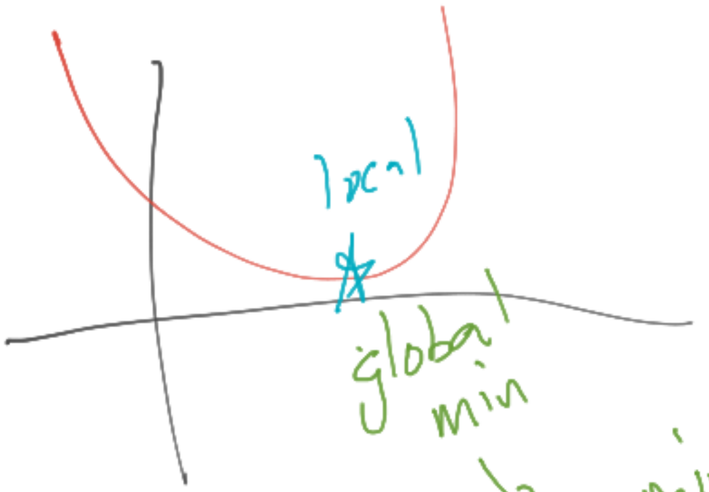


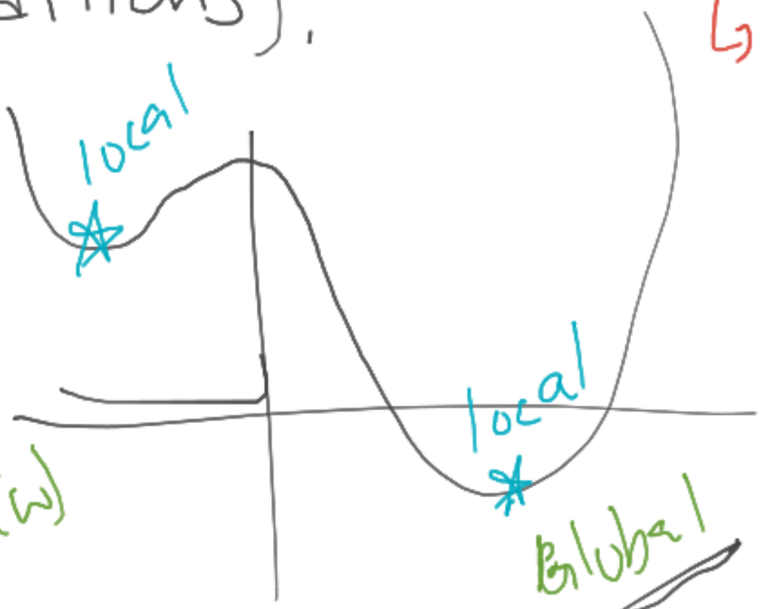
"divergence"

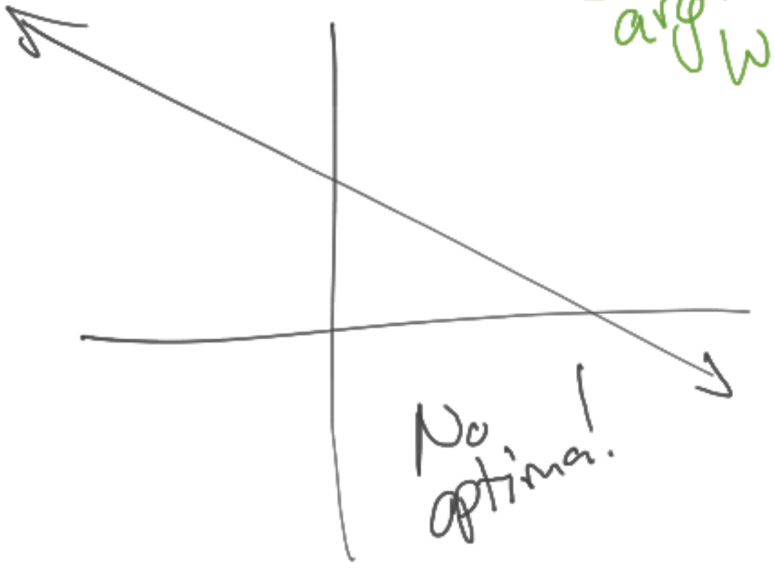

1 week
later.

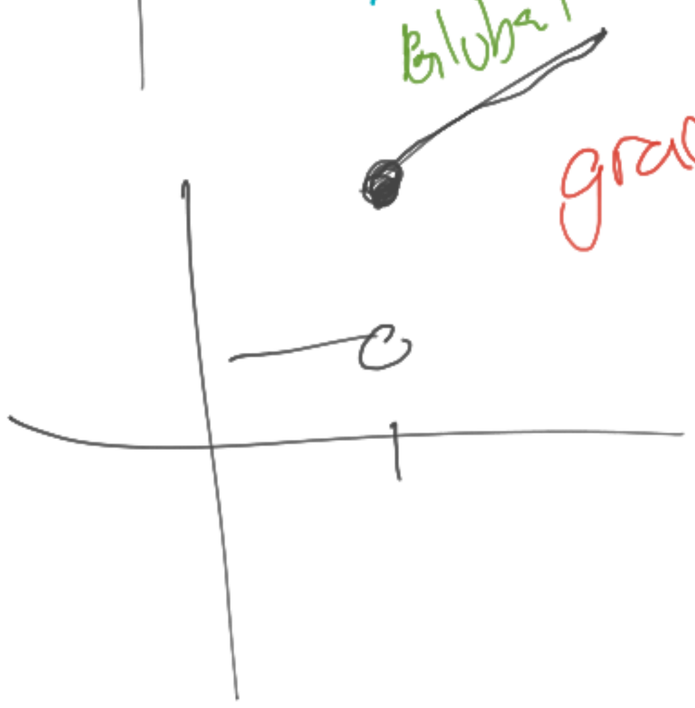
Gradient descent converges to a local minimum (given certain conditions).

↳ no descent directions.



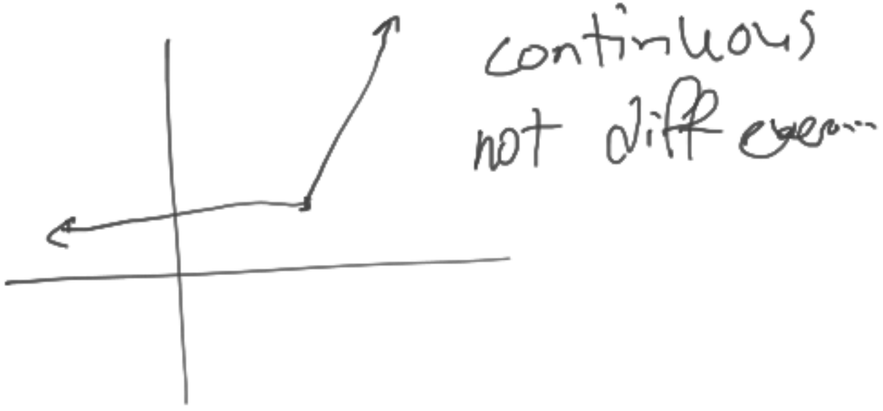






gradient not defined at discontinuity.

- Can diverge if step size too big.
- Requires loss function to be differentiable everywhere.



Linear least squares

$$l(w_i) = \sum_{k=1}^{n} \left(y_k - \sum_{j=1}^{m} w_{i,j} x_{k,j} \right)^2$$

$$cw^2 + c_2 w_{i,1} w_{i2} + c$$

- quadratic.
- One local min, the global min.

Common requirements for convergence:

- ℓ is "smooth" (differentiabe everywhere)
- ℓ is Lipschitz with constant L. → "maximum slope is ℓ"
- Step size is ... appropriately set.
  - → Sufficiently small: $\alpha < \frac{1}{L}$
  - → Decayed appropriately. (square summable but not summable).

$$\sum_{i=0}^{\infty} \alpha_i = \infty$$

$$\sum_{i=0}^{\infty} \alpha_i^2 < \infty$$

- $\nabla \ell$ is Lipschitz.
- ℓ is convex, quadratic, strongly convex.

Types of convergence:

$$w_t \rightarrow \arg\min_{w} \ell(w)$$

- $\nabla \ell(w_i) \rightarrow 0$
- $\nabla \ell(w_i) \rightarrow 0$ <u>OR</u> $\ell(w_i) \rightarrow -\infty$

$$w_{k+1,j} = w_{k,j} - \alpha \boxed{\frac{\partial \ell(w_k)}{\partial w_{k,j}}}$$

$$\ell(w_k) = \frac{1}{2}\sum_{i=0}^{n}\left(y_i - f_{w_k}(x_i)\right)^2$$

$$f_w(x_i) = \sum_{j=1}^{m} w_j x_{ij}$$

$$\frac{\partial}{\partial x} x^2 \to 2x$$

$$\frac{\partial}{\partial x} f(x)^2 \to 2f(x)\frac{\partial f(x)}{\partial x}$$

$$\frac{\partial}{\partial w_{k,j}} \ell(w_k) = \frac{\partial}{\partial w_{k,j}} \frac{1}{2}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2$$

$$= \sum_{i=0}^{n} \frac{\partial}{\partial w_{k,j}} \frac{1}{2}(y_i - \hat{y}_i)^2$$

$$= \sum_{i=0}^{n} \cancel{2}\cdot\frac{1}{\cancel{2}}(y_i - \hat{y}_i)\frac{\partial}{\partial w_{k,j}}(y_i - \hat{y}_i)$$

$$= \sum_{i=0}^{n}(y_i - \hat{y}_i)\frac{\partial}{\partial w_{k,j}}\sum_{\beta=1}^{m} w_{k,\beta} x_{i,\beta}$$

skipping steps!

$$= -\sum_{i=0}^{n}(y_i - \hat{y}_i)\frac{\partial}{\partial w_{k,j}} w_{k,j} x_{i,j}$$

$\frac{\partial}{\partial z} z \cdot \text{const} = \text{const}$

$$= -\sum_{i=0}^{n}(y_i - \hat{y}_i) x_{i,j}$$