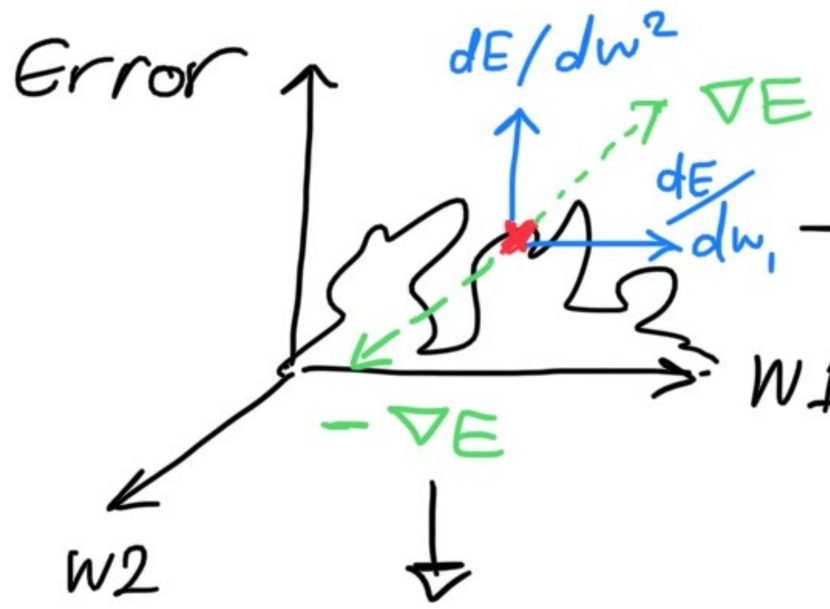
Gradient Descent



→ minimizing the error means reaching axis w2 as much as we can.

-∇E tells us how to decrease E the most.

In models:

$\hat{y} = \sigma(Wx+b)$ Bad

$\hat{y} = \sigma(w_1x_1 + \cdots + x_nx_n + b)$

$\nabla E = \frac{\partial E}{\partial w_1} + \cdots + \frac{\partial E}{\partial w_n} + \frac{\partial E}{\partial b}$, $\alpha = 0.1$ learning rate

$W_i' \leftarrow W_i - \alpha \frac{\partial E}{\partial w_i}$ we want to take small steps.

$b' \leftarrow b - \alpha \frac{\partial E}{\partial w_i}$

$\Longrightarrow \hat{y} = \sigma(W'x + b')$ ← better

replacing

$\Longrightarrow \nabla E = -(y-\hat{y})(x_1, \cdots, x_n, 1)$

E ↳ gradient at point $(x_1, \cdots, x_n)$ label y

$\Longrightarrow w_i' \leftarrow w_i + \alpha(y-\hat{y})x_i$

$b' \leftarrow b + \alpha(y-\hat{y})$