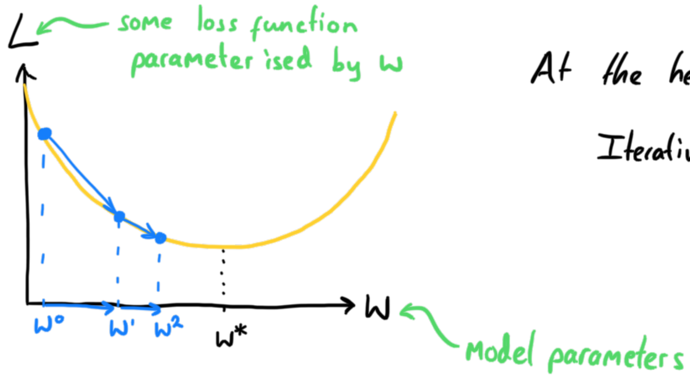
Gradient descent



At the heart of it:

Iteratively move <u>down the slope</u>

follow this heuristic, <u>the gradient</u>

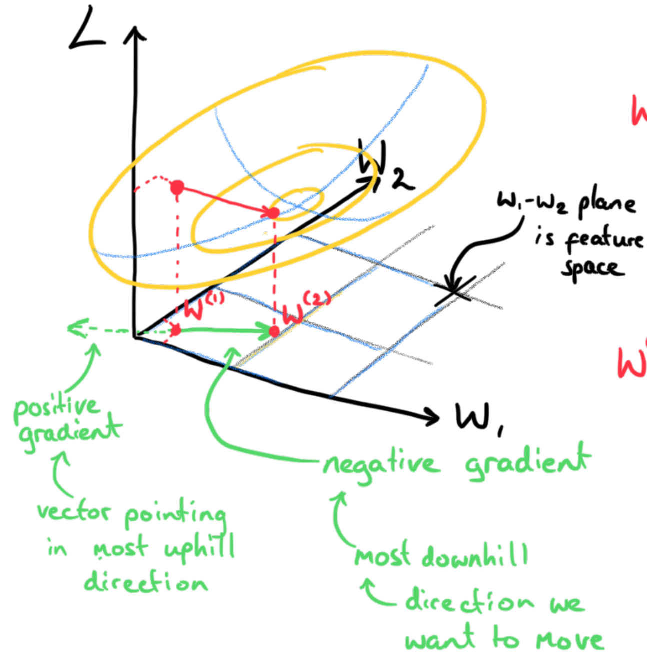
We update the model parameters as follows

$$\boxed{w \leftarrow w - \alpha \frac{\partial L}{\partial w}}$$

gradient descent update rule

"set the new param equal to the old param value shifted in the direction of the negative gradient"

Most models of interest have multiple params, so let's get an idea of how this looks for a <u>higher dimensional feature space</u>



$$w^{(1)} = \begin{bmatrix} w_1^{(1)} \\ w_2^{(1)} \end{bmatrix}$$

$$w^{(2)} = \begin{bmatrix} w_1^{(2)} \\ w_2^{(2)} \end{bmatrix} = w^{(1)} - \alpha \frac{\partial L}{\partial w}\bigg|_{w=}$$

evaluated where w = w

$$= \begin{bmatrix} w_1^{(1)} \\ w_2^{(1)} \end{bmatrix} - \alpha \begin{bmatrix} \frac{\partial L}{\partial w_1} \\ \frac{\partial L}{\partial w_2} \end{bmatrix} = \begin{bmatrix} w_1^{(1)} - \alpha \frac{\partial}{\partial} \\ w_2^{(1)} - \alpha \frac{\partial}{\partial} \end{bmatrix}$$

Notice that the gradient is a vector $\longrightarrow \frac{\partial L}{\partial W}$

it's always the same shape as the model parameters

$\hookrightarrow$ WHY? $\longrightarrow$ because we need an update for each parameter

As long as we can compute $\frac{\partial L}{\partial W}$ then we can do gradient descent to descend the loss function in parameter space & optimise the params

So how does that work in the case of linear regression?

Our model: $\hat{y} = XW + b$

Our loss: $L = \frac{1}{m}\sum_{i=1}^{m}(\hat{y} - y)^2$

$$= \frac{1}{m}\sum^{m}(XW + b - y)^2$$

$$= \mathbb{E}(XW + b - y)^2$$

$$\frac{\partial L}{\partial W} = \frac{\partial}{\partial W}\left(\frac{1}{m}\sum^{m}(XW + b - y)^2\right)$$

$$= \frac{1}{m}\sum^{m}\underbrace{\frac{\partial}{\partial W}\left((XW + b - y)^2\right)}_{= 2(XW + b - y)X}$$

gradient of mean = mean of gradient

$$= \underbrace{\frac{1}{m}\sum^{m}}_{\text{average over each example}} 2\underbrace{\underbrace{\left(x^{(i)}W + b - y^{(i)}\right)}_{\in \mathbb{R}}\underbrace{x^{(i)}}_{\in \mathbb{R}^n}}_{\substack{\in \mathbb{R}^n \\ \text{vector of gradients}}}$$

Can you figure out how to compute this in vector/matrix form?

Same procedure for bias:

$$\frac{\partial L}{\partial b} = \frac{\partial}{\partial b}\left(\frac{1}{m}\sum^{m}(XW+b-y)^2\right)$$

gradient of mean = mean of gradient

$$= \frac{1}{m}\sum^{m}\underbrace{\frac{\partial}{\partial b}\left((XW+b-y)^2\right)}_{=2(XW+b-y)}$$

$$= \underbrace{\frac{1}{m}\sum^{m}}_{\text{average over the vector}} 2\underbrace{(XW+b-y)}_{\in \mathbb{R}^m}$$