Gradient descent

The fundamentals

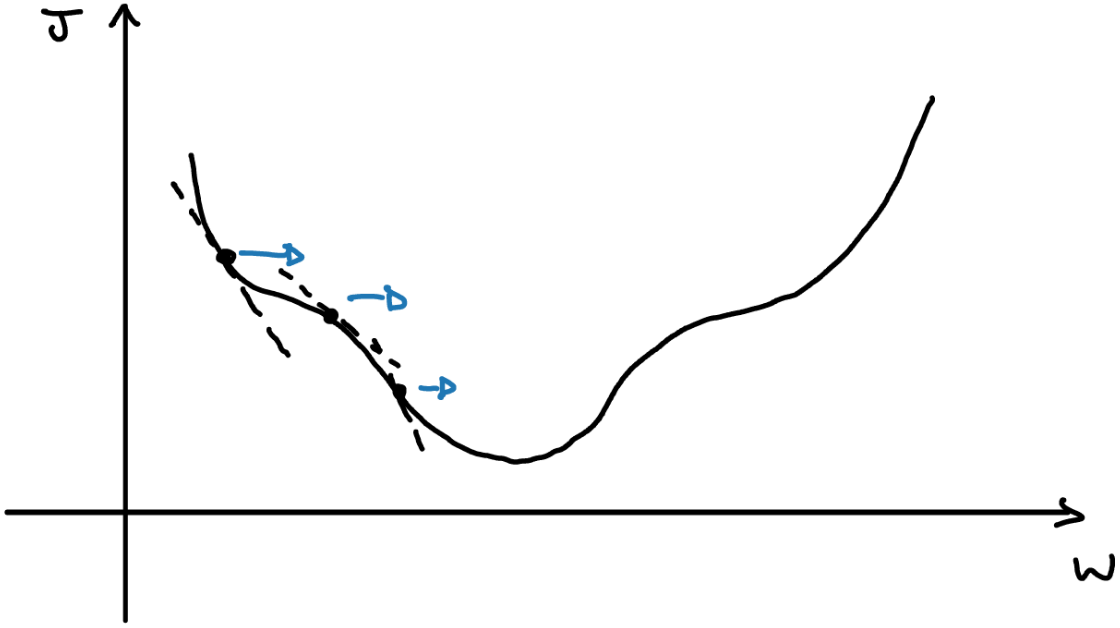
Herman Kamper

http://www.kamperh.com/

Gradient descent

- We have some function $J(\mathbf{w})$ that we want to minimise w.r.t. parameters $\mathbf{w}$
- **Idea:** Start with a random $\mathbf{w}$ and then keep updating it to reduce $J(\mathbf{w})$

In one dimension:

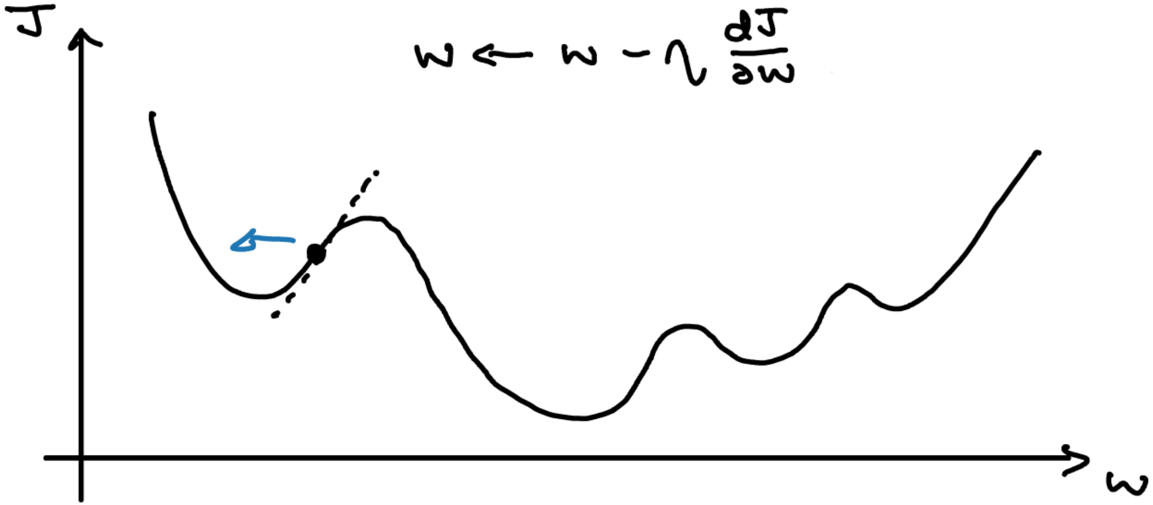


$$w \leftarrow w - \eta \frac{dJ}{dw}$$

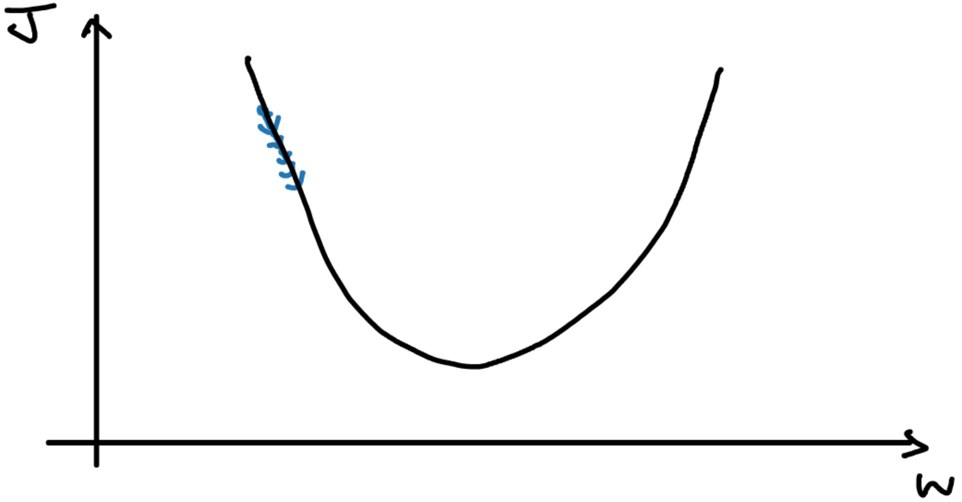
Learning rate

Potential problems:

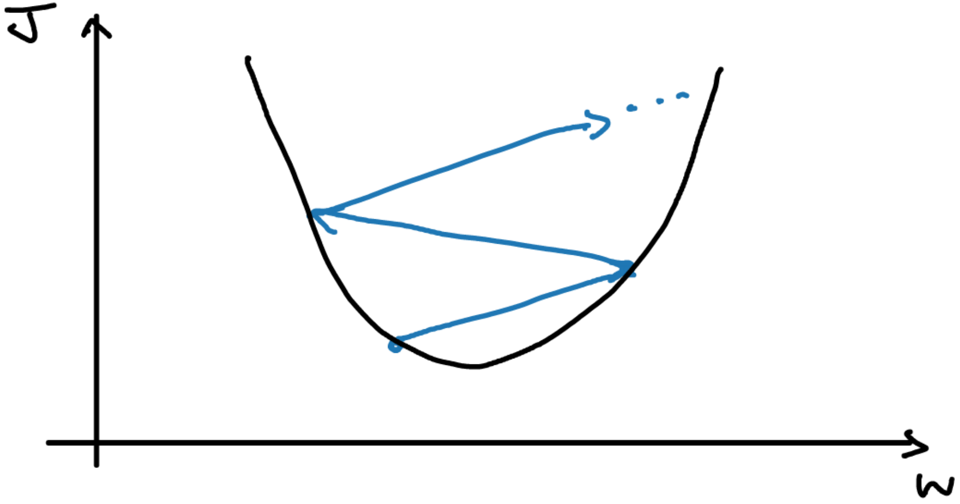
Could get stuck in a local minimum :



If η too small:

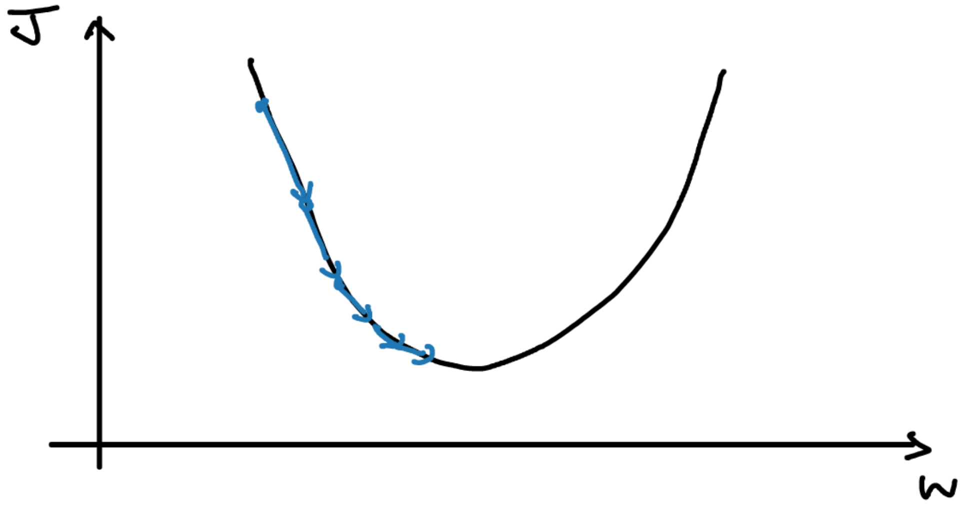


If η too big:



Step sizes

As we get closer to the minimum, the step sizes automatically gets smaller:



In D dimensions

$$w_0 \leftarrow w_0 - \eta \frac{\partial J}{\partial w_0}$$

$$w_1 \leftarrow w_1 - \eta \frac{\partial J}{\partial w_1}$$

$$\vdots$$

$$w_D \leftarrow w_D - \eta \frac{\partial J}{\partial w_D}$$

$$\underline{w} \leftarrow \underline{w} - \eta \frac{\partial J}{\partial \underline{w}}$$

Could even be a matrix