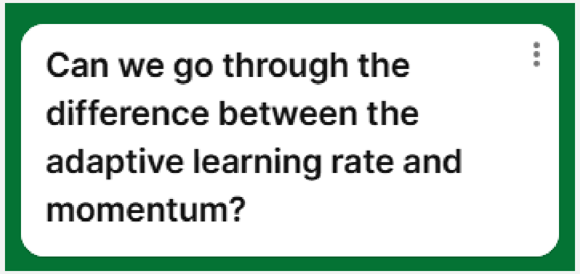
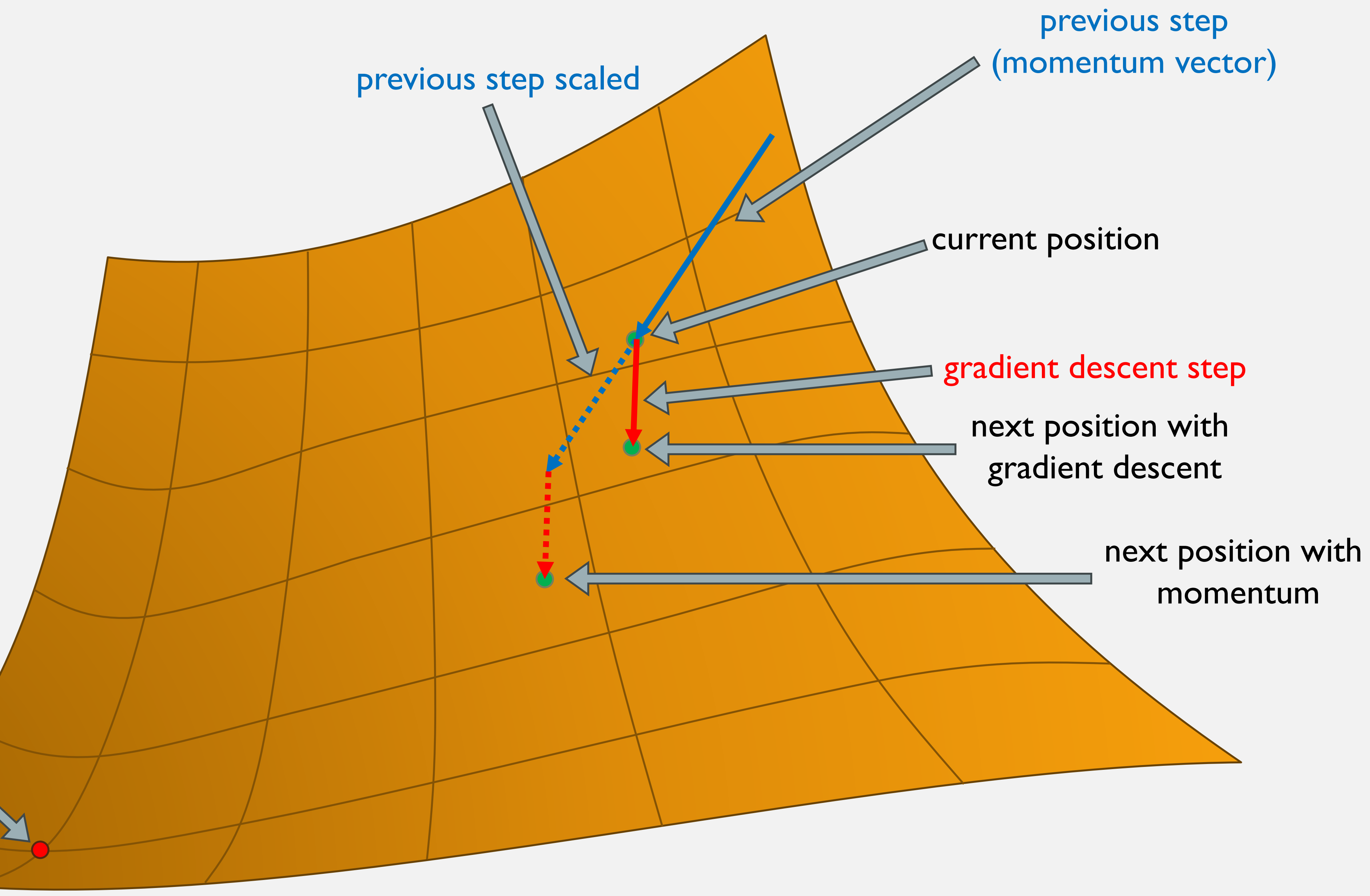
DIVING INTO THE MACHINE ROOM

Pain points



Can we go through the difference between the adaptive learning rate and momentum?



previous step scaled

previous step (momentum vector)

current position

gradient descent step

next position with gradient descent

next position with momentum

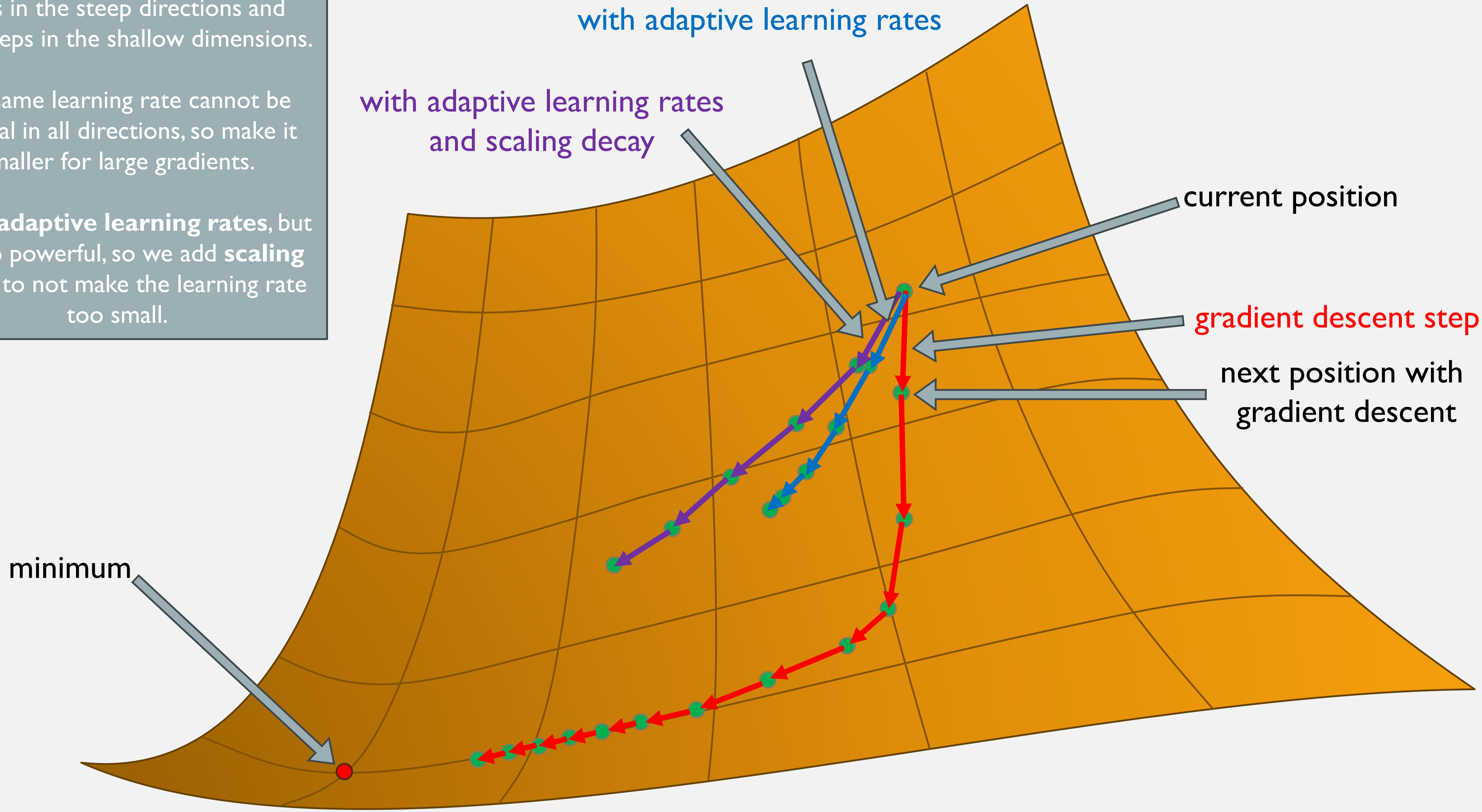
minimum

Momentum: Add a bit of the previous step

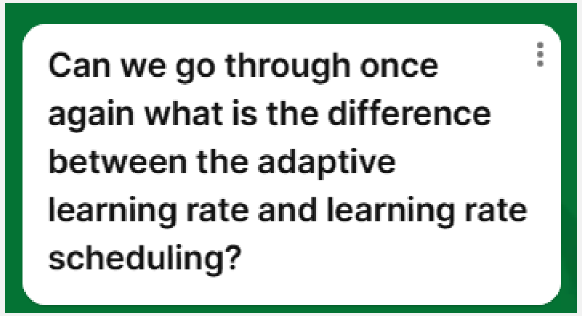
With gradient descent, we take big steps in the steep directions and small steps in the shallow dimensions.

The same learning rate cannot be optimal in all directions, so make it smaller for large gradients.

This is **adaptive learning rates**, but it's too powerful, so we add **scaling decay** to not make the learning rate too small.


with adaptive learning rates
with adaptive learning rates and scaling decay
current position
gradient descent step
next position with gradient descent
minimum

Combine momentum, adaptive learning rates and scaling decay to get Adam.



Can we go through once again what is the difference between the adaptive learning rate and learning rate scheduling?

They solve two different problems!

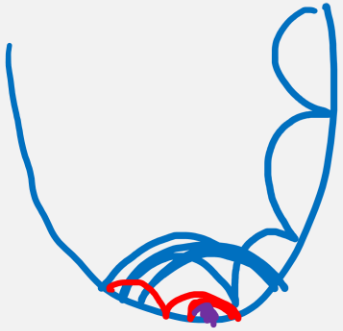
2 different problems

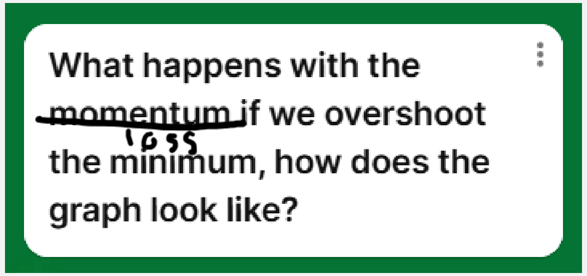
Adaptive LR

Getting η right in several dimensions

LR scheduling

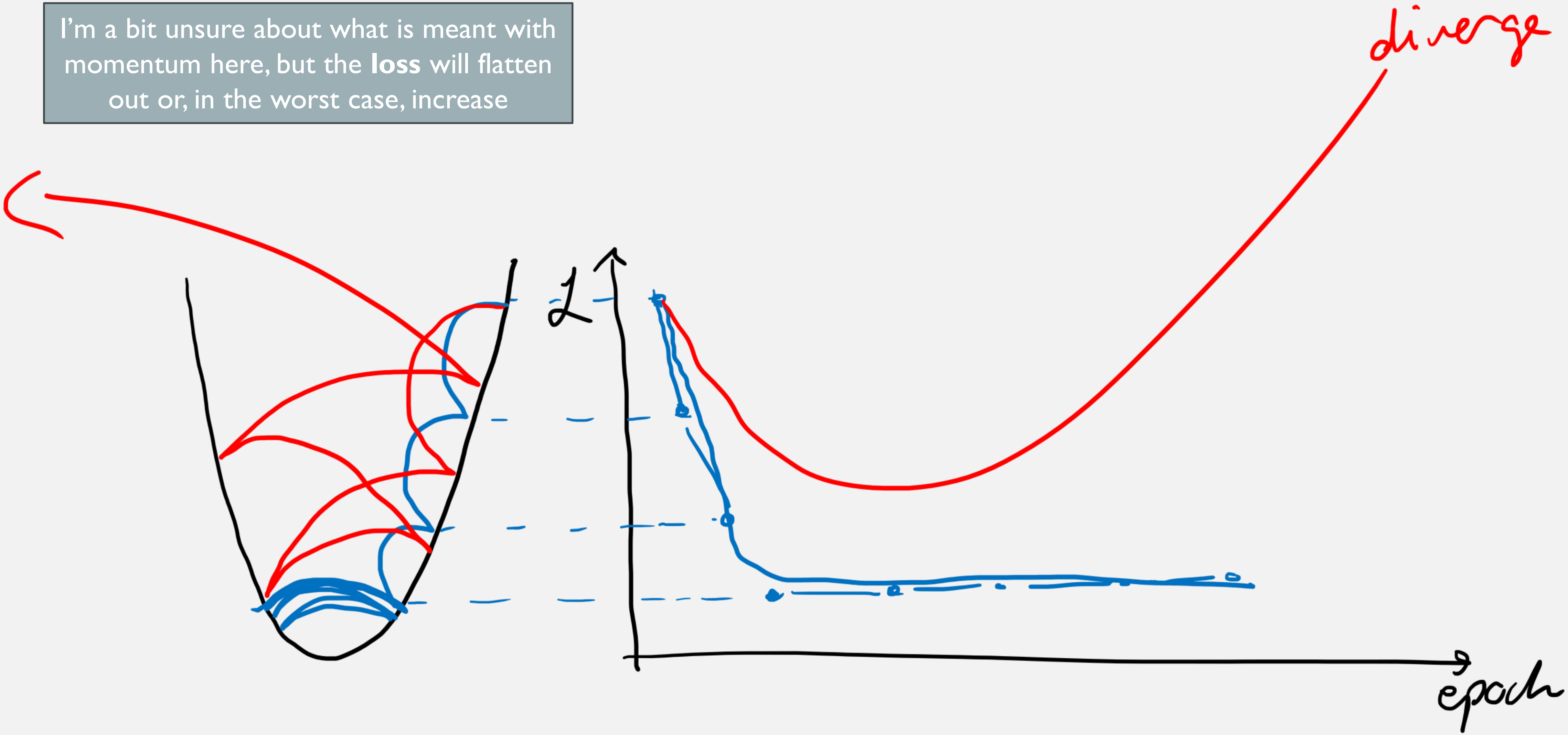
Getting η right as training progresses





What happens with the ~~momentum~~ loss if we overshoot the minimum, how does the graph look like?

I'm a bit unsure about what is meant with momentum here, but the **loss** will flatten out or, in the worst case, increase


diverge
$\mathcal{L}$
epoch