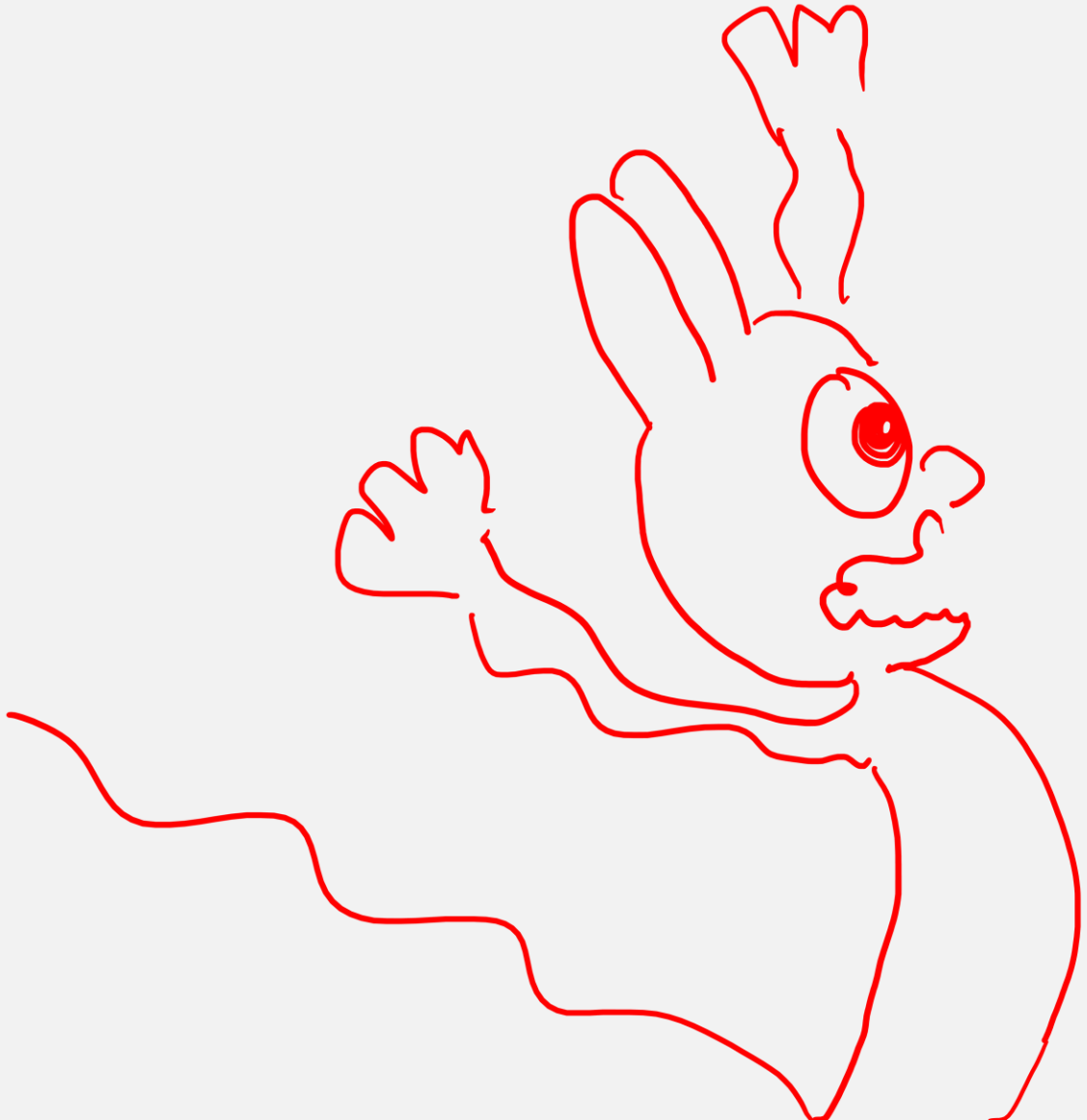
DIVING INTO THE MACHINE ROOM

Pain points

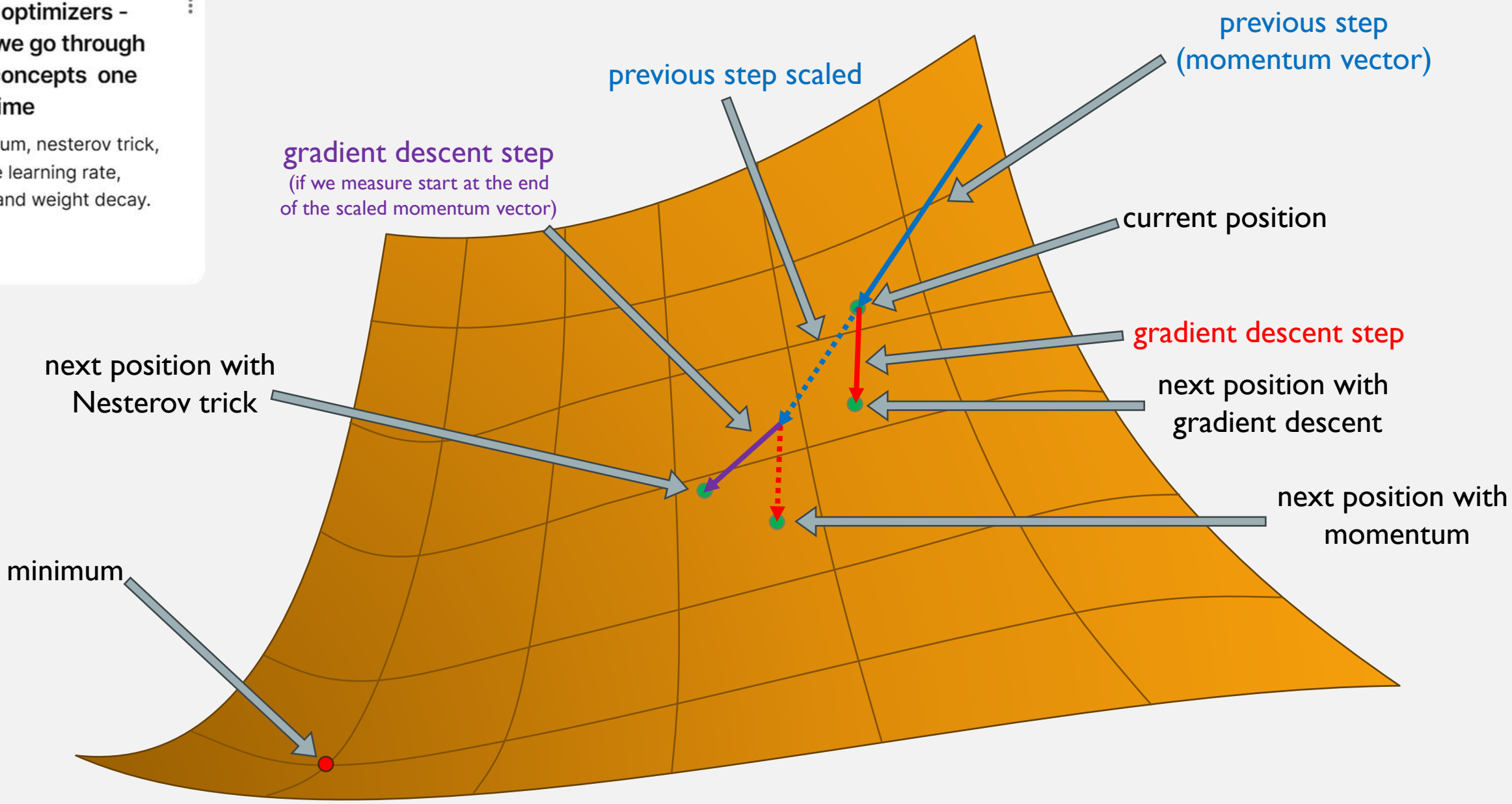
rabbit stew

Explain the analogy with rabbits and fire again please



Faster optimizers - could we go through the 5 concepts one more time

Momentum, nesterov trick, adaptive learning rate, scaling and weight decay.

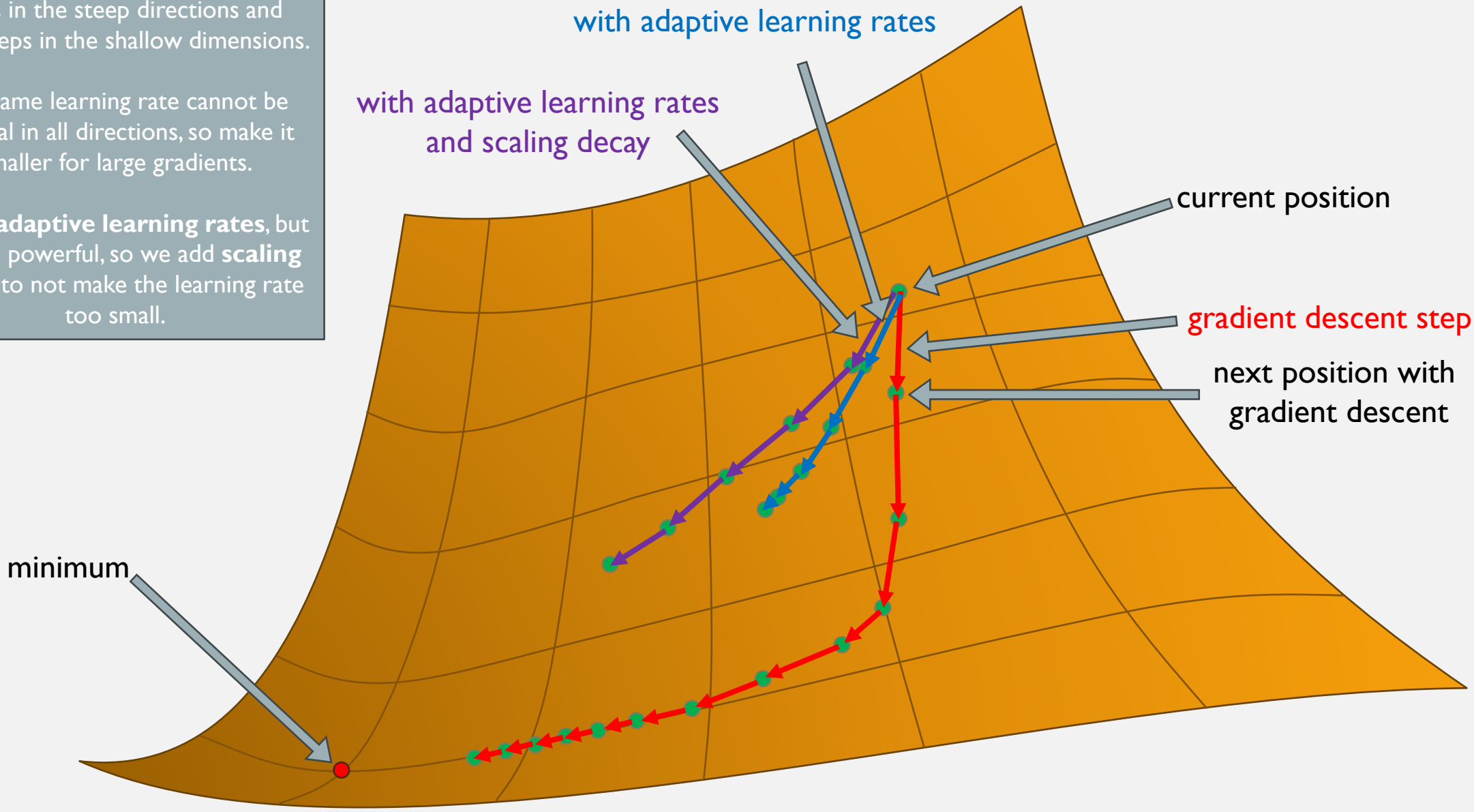


Momentum: Add a bit of the previous step
Nesterov: Measure the gradient *after* the momentum step

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.



Combine momentum, adaptive learning rates and scaling decay to get Adam.
Add Nesterov to get Nadam.
Add weight decay (for regularization) to get AdamW.