The simplicity of gradient descent (GD) made it the default method for training ever-deeper and complex neural networks. Both loss functions and architectures