<https://blog.paperspace.com/intro-to-optimization-in-deep-learning-gradient-descent/>

<https://blog.paperspace.com/intro-to-optimization-momentum-rmsprop-adam/>

A widely used technique in gradient descent is to have a variable learning rate, rather than a fixed one. Initially, we can afford a large learning rate. But later on, we want to slow down as we approach a minima. An approach that implements this strategy is called **Simulated annealing**, or decaying learning rate.

In stochastic gradient descent, instead of taking a step by computing the gradient of the loss function creating by summing all the loss functions, we take a step by computing the gradient of the loss of only one randomly sampled (without replacement) example. In contrast to **Stochastic Gradient Descent**, where each example is stochastically chosen, our earlier approach processed all examples in one single batch, and therefore, is known as **Batch Gradient Descent.**

This means, at every step, we are taking the gradient of a loss function, which is different from our actual loss function (which is a summation of loss of every example). The gradient of this "one-example-loss" at a particular may actually point in a direction slighly different to the gradient of "all-example-loss".