some notes..

(Mometum)

- It differs a little from SGD such that the movement through parameter space is averaged over multiple steps. It does so by introducting velocity component in the direction of large improvements; helping the network avoids local minima.

- alpha: momentum parameter: controls how fast the velocity can change and how much local grad influences long-term movements.

(Nestrov Momentum)

- Unlike Normal Momentum, the grad is NOT calc from the current position, but from an intermediate position, which is helpful! Having the grad always pointing at the right dir, the momentum might not!
While the momentum might overshoots aiming in the wrong dire, the grad can still go back & correct it in the same update step.

- Both Momentum & Nestrov Momentum handle such complex fn.
Both are sensitive to learning rate.