Adam: Adam will compute the adaptive learning rates for each parameter. It keeps the exponentially decaying average of past squared gradients and gradients . and are the estimates of first and second moment of gradient as bias. It performance with low training loss and error with high accuracy. However, Adam use to adjust learning rate which will cause instability.

其中，表示当前步的梯度，表示在当前步的更新后得到的参数，是学习率，和是滑动平均系数，和分别是梯度和梯度平方的指数移动平均值，是一个很小的数，用于防止除零错误。和是对和进行偏差校正后得到的估计值，避免了在训练初期梯度平均值和梯度平方平均值的偏差较大的情况。

Nadam: Nadam is an extension of Adam, it incorporates with N(Nesterov) acceleration to compute the gradient in the next step, which can help speed up convergence. Also, Nadam introduces extra bias-correction step to reduce oscillations during training. In the graph, it lead to similar optimization performance compared to Adam but with better convergence and robustness.

其中，表示当前步的梯度，表示在当前步的更新后得到的参数，是学习率，和是滑动平均系数，和分别是梯度和梯度平方的指数移动平均值，是一个很小的数，用于防止除零错误。和是对和进行偏差校正后得到的估计值。是通过使用Nesterov加速方式得到的中间值，可以更准确地估计下一步的梯度。

AMSGard: AMSGard retains the maximum of all past second moment estimates to prevent sudden drops in learning rate, resulting in better optimization performance. Also, AMSGard uses a correction factor to ensure the gradient average is not underestimated during the iteration in order to get the better convergence performance. However, AMSGard need more time to compute.

其中，表示当前步的梯度，表示在当前步的更新后得到的参数，是损失函数，是学习率，和是滑动平均系数，和分别是梯度和梯度平方的指数移动平均值，是一个很小的数，用于防止除零错误。是计算出来的归一化梯度，用于调整更新步长，使得不同的参数在更新时具有相似的步长。

SGD: SGD only compute the loss function according to the gradient to update the parameter, so the compute time of SGD is smaller than other optimizers. However, it doesn’t have dynamic adjusting of learning rate, so the performance of SGD is worse than other adaptive optimization algorithms.

其中，表示当前步的梯度，表示在当前步的更新后得到的参数，是损失函数，是学习率，和是训练数据中的输入和输出。