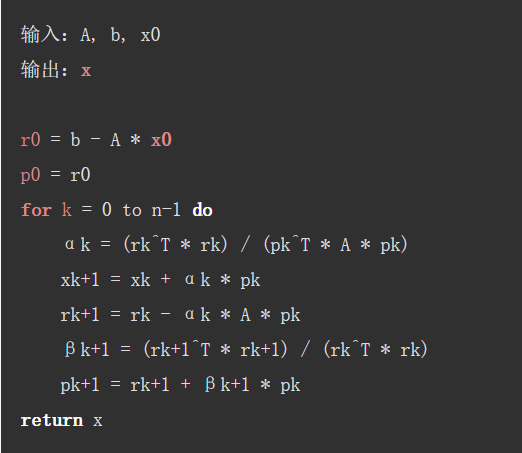
## 调研报告：常用的优化算法

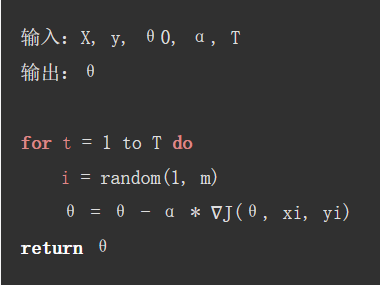
1. 引言

在深度学习中，模型的优化方法是至关重要的。与传统机器学习中仅仅使用梯度下降法不同，深度学习中有很多种优化算法，如共轭梯度法、随机梯度下降法SGD、AdapGrad、动量法（Momentum）和Adam算法等。这些算法的引入，使得深度学习的优化变得更加高效和精确。本文将对这些算法进行介绍和比较。

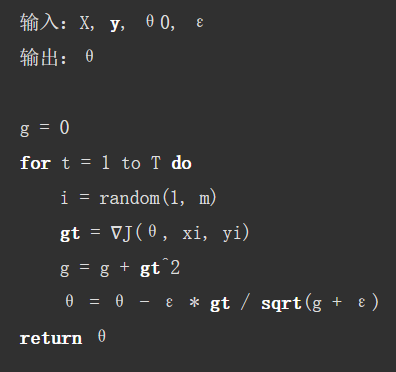
1. 共轭梯度法：
2. 发明人和时间：共轭梯度法是由Hestenes和Stiefel在1952年提出的。
3. 背景和对应的论文：共轭梯度法是为了解决线性方程组的求解问题而提出的，但后来也被应用于非线性优化问题。Hestenes和Stiefel在1952年的论文《Methods of conjugate gradients for solving linear systems》中详细介绍了该算法。
4. 主要思路：共轭梯度法是一种迭代算法，用于求解线性方程组或最小化二次函数。它利用了线性方程组的系数矩阵的正定性质，以加速收敛速度。该算法的主要思路是在每次迭代中，利用上一步的搜索方向和梯度信息，计算出一个新的搜索方向，并沿着这个方向更新解的值。
5. 主要步骤（伪代码）：



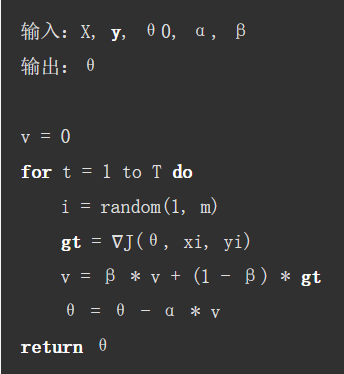
1. 优缺点：优点： - 收敛速度快，通常比梯度下降法要快得多。 - 可以处理大规模的线性方程组。 - 可以利用前一次迭代的信息，减少计算量。 缺点： - 只能用于求解线性方程组或最小化二次函数。 - 对于非正定的矩阵，可能会出现振荡现象。
2. 随机梯度下降法SGD：
3. 发明人和时间：随机梯度下降法SGD是由Robbins和Monro在1951年提出的。
4. 背景和对应的论文：随机梯度下降法SGD是为了解决大规模数据集的优化问题而提出的。Robbins和Monro在1951年的论文《A stochastic approximation method》中详细介绍了该算法。
5. 主要思路：随机梯度下降法SGD是一种基于梯度的优化算法，用于最小化目标函数。与传统的梯度下降法不同，它每次只随机选择一个样本计算梯度，并沿着该方向更新模型参数。由于每次只计算一个样本的梯度，因此可以处理大规模的数据集，同时也可以避免陷入局部最优解。
6. 主要步骤（伪代码）：



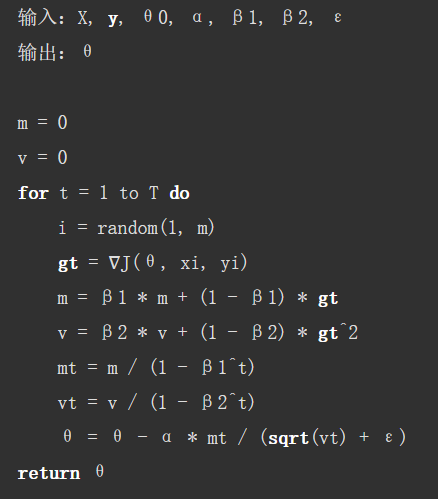
1. 优缺点：优点： - 可以处理大规模的数据集。 - 可以避免陷入局部最优解。 - 计算速度快。 缺点： - 更新方向是随机的，可能会出现震荡现象。 - 对于非凸函数，可能会收敛到次优解。
2. AdapGrad：
3. 发明人和时间：AdapGrad是由Duchi、Hazan和Singer在2011年提出的。
4. 背景和对应的论文：AdapGrad是为了解决梯度下降法中学习率的调整问题而提出的。Duchi、Hazan和Singer在2011年的论文《Adaptive subgradient methods for online learning and stochastic optimization》中详细介绍了该算法。
5. 主要思路 AdapGrad是一种自适应学习率的优化算法，用于最小化目标函数。它通过对每个参数的学习率进行自适应调整，使得每个参数的学习率与其历史梯度的平方根成反比。这样可以使得较少更新的参数具有较大的学习率，而较频繁更新的参数具有较小的学习率，从而在训练过程中更加稳定。
6. 主要步骤（伪代码）：



1. 优缺点：优点： - 自适应学习率，可以提高训练的稳定性。 - 可以处理非稳态的训练数据。 缺点： - 可能会出现学习率过小的情况，导致收敛速度慢。
2. 动量法(Momentum) ：
3. 发明人和时间：动量法(Momentum)是由Polyak在1964年提出的。
4. 背景和对应的论文 动量法(Momentum)是为了解决梯度下降法中收敛速度慢的问题而提出的。Polyak在1964年的论文《Some methods of speeding up the convergence of iteration methods》中详细介绍了该算法。
5. 主要思路 动量法(Momentum)是一种基于梯度的优化算法，用于最小化目标函数。与传统的梯度下降法不同，它引入了一个动量项，用于加速收敛速度。具体来说，它利用了之前的梯度信息，计算出一个动量，然后将其与当前的梯度方向相结合，更新模型参数。
6. 主要步骤（伪代码）：



1. 优缺点：优点： - 可以加速收敛速度。 - 可以处理非凸函数。 缺点： - 可能会出现震荡现象。 - 对于某些问题，可能会收敛到次优解。
2. Adam算法：
3. 发明人和时间 Adam算法是由Kingma和Ba在2015年提出的。
4. 背景和对应的论文 Adam算法是为了解决梯度下降法中学习率的调整问题和动量法中可能出现的震荡现象而提出的。Kingma和Ba在2015年的论文《Adam: A Method for Stochastic Optimization》中详细介绍了该算法。
5. 主要思路 Adam算法是一种自适应学习率和动量的优化算法，用于最小化目标函数。它利用了梯度的一阶矩估计和二阶矩估计，对学习率进行自适应调整，并引入了动量项，用于加速收敛速度。具体来说，它计算出每个参数的梯度的一阶矩估计和二阶矩估计，并根据这些估计值计算出一个自适应的学习率和动量。
6. 主要步骤（伪代码）：



1. 优缺点：- 可以自适应地调整学习率和动量。 - 可以处理大规模的数据集。 - 可以处理非凸函数。 缺点： - 对于某些问题，可能会收敛到次优解。
2. 总结：

本文介绍了共轭梯度法、随机梯度下降法SGD、AdapGrad、动量法(Momentum)和Adam算法这五种常用的优化算法。这些算法各有优缺点，在实际应用中需要根据具体情况选择合适的算法。

参考网页[[1]](#endnote-0)

1. [https://blog.csdn.net/wyz6666/article/details/98885164](https://blog.csdn.net/wyz6666/article/details/98885164；)

   <https://blog.csdn.net/Scarlett_Guan/article/details/98957906>

   <https://blog.csdn.net/zhenjiteng/article/details/128168348>

   <https://link.springer.com/chapter/10.1007/BFb0055275> [↑](#endnote-ref-0)