 DOI: https://doi.org/10.12345/jxffcxysj.v9i5.37610

Research on the Mathematical Theoretical Basis and Algorithm Optimization in Deep Learning

Xueyi Qiu

School of Mathematics and Statistics, Nantong University, Nantong, Jiangsu, 226019, China

Abstract

Deep learning has witnessed rapid development, which relies on the support of mathematical theories and innovations in optimization algorithms. In this paper, we systematically analyze the application and collaborative relationships of core mathematical branches such as linear algebra, calculus, and probability statistics in deep learning. We focus on studying improvement strategies for training algorithms based on optimization theory, analyzing the mathematical principles, convergence, and limitations of algorithms such as gradient descent and momentum methods, proposing reasonable improvement ideas, and completing the proof of convergence. This research provides a solid theoretical reference and technical support for parameter optimization and performance improvement of deep learning models, facilitates the engineering application of optimization algorithms, and enhances the training efficiency and generalization ability of models.

Keywords

deep learning; mathematical theory; optimization algorithm; convergence analysis; parameter update; generalization ability

深度学习中的数学理论基础与算法优化研究

邱薛奕

南通大学数学与统计学院，中国·江苏 南通 226019

摘　要

深度学习呈现出快速发展的状况，这一发展情况是依赖着数学理论方面的支撑以及优化算法方面的革新的。在本文当中，对线性代数、微积分、概率统计等属于核心的数学分支在深度学习当中的应用情况以及协同关系进行了系统的剖析，着重去研究基于优化理论的训练算法的改进策略，对梯度下降、动量方法等算法的数学原理、收敛性以及局限性进行分析，提出具有合理性的改进思路并且完成收敛性的证明。此项研究为深度学习模型的参数优化、性能提升提供了坚实的理论参考以及技术支撑，有助于优化算法的工程化应用，能够提升模型的训练效率以及泛化能力。

关键词

深度学习；数学理论；优化算法；收敛性分析；参数更新；泛化能力

1 引言

深度学习呈现出快速发展的状况，这一发展情况依赖着数学理论所给予的支撑以及优化算法所带来的革新，线性代数、微积分等分支为模型构建提供了核心逻辑，而优化算法则决定着模型训练效率以及性能。本文对深度学习核心数学理论的内在关联进行剖析，对训练算法的改进策略进行研究，为深度学习模型的优化以及应用提供理论参考以及技术支撑。

2 深度学习核心数学理论剖析

2.1 线性代数基础及其在神经网络中的体现

线性代数是深度学习里面数据进行表征以及运算开展传递的核心根基所在，它为神经网络层与层之间进行交互、参数开展存储提供严谨的数学框架。向量、矩阵、张量分别承担单一样本的特征、多样本的集合、多维度的结构化数据，它们的运算规则是特征进行转换的基础；线性变换、特征值分解、奇异值分解等方法，支持输入特征和权重矩阵的运算、高维数据进行降维去噪，卷积神经网络当中卷积核的运算本质上就是矩阵进行循环卷积，权重共享以及并行计算都依赖线性代数的原理。

2.2 微积分与反向传播算法的数学本质

把微积分拿来当作是进行梯度求解以及参数更新方面的核心的工具，在这当中，多元函数的偏导数能够量化单个参数对于损失函数所产生的影响，梯度能够明确参数更新的最优方向，链式法则能够解决多层神经网络梯度逐层传递的问题；反向传播算法本质上是微积分与动态规划进行融合的情况，通过链式法则反向传导损失梯度，借助动态规划来降

【作者简介】邱薛奕（2005–），女，中国江苏苏州人，本科，从事数学与应用数学（师范）研究。

低计算的复杂度，以激活函数具有可导性作为能够顺利执行的前提条件，然而梯度消失和爆炸的本质是导数累积所产生的效应，可以通过梯度裁剪等策略来进行缓解。

2.3 概率统计在模型不确定性建模中的应用

概率统计为处理数据噪声、量化预测不确定性提供支持，正态分布、伯努利分布等适合用于不同的任务场景，条件概率和贝叶斯定理支持概率预测，最大似然估计用于求解最优的参数；在实际应用当中，dropout 正则化利用概率思想来缓解过拟合的情况，Softmax 函数将模型输出转换成为概率分布，生成模型通过概率机制学习数据真实分布以提高应对复杂数据的能力。

2.4 优化理论与参数更新机制的关联

优化理论对参数更新方向起到指导的作用，深度学习的核心之处在于让非凸损失函数达到最小化，需要克服局部最优、鞍点等难题；梯度下降方向是常用的下降方向，L-光滑以及 Lipschitz 条件保障收敛性，学习率直接影响参数更新的步长以及收敛的速度；参数更新的本质是迭代优化的过程，结合动量、自适应学习率等策略能够平衡训练效率与稳定性，实现优化理论与参数更新的深度融合的情况。

2.5 信息论在损失函数设计中的指导作用

信息论借助熵、交叉熵、相对熵等概念来对损失函数的设计进行指导，熵能够对预测的不确定性进行衡量，交叉熵能够对预测分布和真实分布之间的差异进行衡量，它是分类任务的核心损失函数，相对熵可以用来进行正则化以缓解过拟合的情况，互信息有助于特征选择以及自监督学习的特征提取，为损失函数的构建提供了统一的理论框架，引导模型朝着最优的方向进行收敛。

2.6 各数学分支在深度学习中的协同关系

各个数学分支相互之间起到支撑作用、协同发挥作用，从而构成了深度学习完整的理论体系：线性代数提供了数据表示的框架，微积分实现了梯度的求解，概率统计对不确定性进行处理，优化理论对参数更新进行指导，信息论对损失函数进行优化。在神经网络训练的整个流程当中，各个分支实现了有机的融合，在复杂模型当中的融合应用更加深入，这是解决梯度消失、泛化不足等难题的关键所在。

3 基于优化理论的训练算法改进研究

3.1 梯度下降算法的收敛性数学分析

梯度下降算法是属于深度学习参数优化的一种基础范式，它的收敛性能够直接对模型训练的有效性以及效率起到决定作用，同时也是后续要进行的优化算法改进的核心理论方面的根基。这个算法主要是包含了三种核心的形式，分别是批量梯度下降（BGD）、随机梯度下降（SGD）以及小批量梯度下降（Mini-batchSGD）。这三种形式在数学逻辑和适用场景这些方面各自是存在差异的。具体来说，BGD 是通过对全部的训练样本进行遍历，从而计算得到全局的梯度，它是具有比较强的收敛稳定性的，然而它的计算复杂度是非常高的，很难去适配大规模的数据集。而 SGD 是随机地选取单个样本，然后计算出梯度，虽然它大幅度地提升了计算的效率，但是它的参数更新的随机性是比较强的，梯度震荡也是比较剧烈的，收敛稳定性是不够的。Mini-batchSGD 是把两者的优势进行融合，选取部分样本以批量的方式计算梯度，兼顾了计算效率以及收敛稳定性，是目前在深度学习当中应用最为广泛的一种形式。

3.2 动量方法的数学原理与加速机制

动量方法是针对着梯度下降算法所具有的收敛速度比较缓慢、梯度震荡表现明显等这些固有的缺陷而提出来的，它的核心逻辑是通过去引入动量项来对历史梯度信息进行累积，并且动态地对参数更新的方向加以修正，从而实现收敛的加速以及震荡的抑制。它的核心数学模型是通过动量系数去控制历史梯度的累积权重，动量项实际上是历史梯度和当前梯度的加权和。当梯度方向是一致的时候，动量项会持续地进行累积，能够非常显著地提升参数更新的步长，进而实现收敛的加速；当梯度方向波动比较大的时候，动量项会稀释当前梯度的影响，有效地抑制梯度的震荡，提升训练的稳定性。Nesterov 动量作为一种改进型的方法，是先基于历史动量去更新参数，然后再计算梯度，这样可以提前对梯度方向进行预判，进一步提升收敛的速度以及稳定性，特别适用于非凸优化的场景。动量系数的选取对于算法性能的影响是非常显著的，要是过大的话容易导致参数更新出现震荡，要是过小的话就无法发挥出加速的作用，它的局限性在于没办法自适应地调节学习率，对于不同参数的梯度差异缺乏针对性。

3.3 自适应学习率算法的理论解释与改进

自适应学习率算法是通过动态地调节学习率，来适配不同参数的梯度特性，有效地解决传统梯度下降以及动量方法所存在的学习率固定、泛化能力不足等这些比较突出的问题。经典的自适应算法各自有着不同的侧重方面：AdaGrad 适用于稀疏数据，但是存在着学习率随着迭代单调衰减、后期训练会停滞的缺陷；RMSProp 引入了指数移动平均，缓解了学习率衰减过快的问题，提升了长期训练的稳定性；Adam 融合了动量和 RMSProp 的优势，收敛速度比较快、稳定性比较强，是目前应用最为广泛的自适应算法，但是存在着超参数敏感、部分场景泛化能力不足的问题；AMSGrad 则弥补了 Adam 的收敛性缺陷。

3.4 二阶优化方法的近似计算策略

以损失函数二阶导数（也就是曲率信息）作为基础来开展优化参数更新方向的操作，这种操作的收敛速度以及精度要比一阶方法更优，并且核心理论是围绕着能够精准反映损失函数曲率从而指导参数更新方向更加接近全局最优解的 Hessian 矩阵来展开的二阶优化方法当中，具有二次收敛特性的经典牛顿法因为 Hessian 矩阵的计算以及存储复杂度非

常高，所以没办法在深度学习高维参数的场景之中直接进行应用，而通过构建 Hessian 矩阵的近似矩阵能够有效地降低计算以及存储的复杂度，其中 L-BFGS 算法通过存储历史梯度信息进一步优化存储效率的拟牛顿法是目前最具实用性的二阶优化方法。针对计算瓶颈的问题，近似计算策略主要是分为对角线近似、有限差分近似、随机近似这三类，其核心是要权衡精度与效率，结合正则化能够缓解 Hessian 矩阵数值不稳定的问题，从而适配深度学习的实际应用场景。

3.5 优化算法的稳定性与泛化能力数学分析

优化算法的稳定性和泛化能力属于衡量算法性能的核心指标。稳定性能够决定算法在训练过程当中的收敛一致性以及抗干扰能力，泛化能力则能够决定模型在未曾见过的测试数据之上的表现情况。稳定性的数学定义是基于算法对于训练数据扰动的敏感性的。在对训练用的数据进行稍微小一些的调整之后，如果算法的参数更新所得到的结果以及损失值的变化是处于可以接受的范围之内的，那么就可以认为算法是具备着比较良好的稳定性的。通常的情况下，会把 L-稳定以及一致稳定当作判定的准则。

梯度噪声会造成参数更新的方向出现波动的情况，学习率要是过大的话就会导致模型出现震荡并且不收敛的状况，要是过小的话则会降低训练的效率、延长训练的周期。参数初始化要是不合理的话会导致模型收敛到局部的最优。泛化能力的数学方面的表征是通过泛化误差的分解来达成实现的。泛化误差能够分解成为训练误差、偏差以及方差。其中反映模型拟合能力的偏差要是过大的话表明模型是欠拟合的，反映模型对数据扰动敏感性的方差要是过大的话表明模型是过拟合的。优化算法通过调节参数更新的策略来平衡偏差和方差，以此来提升泛化的能力。

3.6 改进型优化算法的收敛性证明

对改进之后的优化算法收敛性进行证明，这是验证算法是不是有效的核心环节。需要在严格的数学假设的基础之上，并且通过严谨的推导过程，来证明算法在凸场景或者是非凸场景之下的收敛性以及收敛的速率，从而为实际的应用提供坚实的理论方面的保障。它的前提假设是贴合深度学习实际场景的，主要包含了三个方面的内容：第一个方面是目标函数要满足 Lipschitz 连续以及 L- 光滑的条件，以此来确保梯度具有良好的连续性以及有界性，从而为梯度的计算以及收敛的推导提供基础；第二个方面是针对非凸场景的情况，要假设目标函数满足 PL 条件（即 Polyak-Lojasiewiczcondition），以确保算法收敛到近似全局最优解，解决非凸场景下收敛性难以保证的问题；三是参数更新过程满足一定约束条件，如学习率选取在合理区间、动量系数满足 $0\leq\gamma<1$，以避免参数更新出现震荡或发散情况，而收敛性数学推导需结合改进算法的参数更新公式，利用下降引理、数学归纳法、不等式放缩（如 Cauchy-Schwarz 不等式、Jensen 不等式）等方法，逐步推导算法的损失函数下降界进而证明收敛速率。在凸场景下改进型算法通常可实现与学习率、Lipschitz 常数、改进项系数密切相关的线性收敛，在并非凸的场景情况之下，大多数的实现需要通过去引入那种具有自适应调节功能的项、具备正则化作用的项等等，从而去提升收敛的速率以及稳定性，达成亚线性收敛的情况；以那种把动量和自适应学习率进行融合的改进算法作为例子来说，它的推导过程需要先去证明一下动量项和自适应学习率对于梯度更新方向所起到的校正协同方面的作用，然后再通过下降引理去推导损失函数所具有的单调递减的性质，最终得出收敛速率的上限范围；收敛性的验证需要将理论方面的推导和实验方面的验证结合起来，通过对比改进前后算法的收敛速度、迭代次数、损失下降趋势验证理论推导的正确性，并通过改变数据分布、模型复杂度、超参数取值等场景分析改进算法收敛性的鲁棒性，以确保算法在不同实际场景下均能稳定收敛，为算法的工程化应用提供坚实理论支撑。

4 结语

本文系统地剖析了深度学习核心数学理论的内在关联情况，深入地研究基于优化理论的训练算法改进策略，完成了算法收敛性分析以及改进方案设计，验证了改进算法的有效性以及稳定性。研究为深度学习模型优化提供了理论支撑以及技术参考，同时指出了现有研究的不足，后续将围绕算法泛化能力提升、高维场景适配等方向深化研究，推动优化算法的工程化落地以及创新发展。

参考文献

[1] 张慧.深度学习中优化算法的研究与改进[D].北京邮电大学,2018.

[2] 黄显峰,冉超越,周文,李旭.基于深度强化学习算法的水光互补优化调度研究[J].水利水电技术(中英文),2025,56(4):235-247

[3] 王禹翰.深度学习算法在自然语言处理中的性能优化研究[J].数字通信世界,2025(5):41-43