计算机科学
COMPUTER SCIENCE

# 基于AdaGrad+的自适应Heavy-Ball动量法及其最优个体收敛性

韦洪旭, 陇盛, 陶蔚, 陶卿

引用本文

韦洪旭, 陇盛, 陶蔚, 陶卿. 基于AdaGrad+的自适应Heavy-Ball动量法及其最优个体收敛性[J]. 计算机科学, 2023, 50(11): 220-226.

WEI Hongxu, LONG Sheng, TAO Wei, TAO Qing. Adaptive Heavy-Ball Momentum Method Based on AdaGrad+ and Its Optimal Individual Convergence [J]. Computer Science, 2023, 50(11): 220-226.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900131

# 基于 AdaGrad+的自适应 Heavy-Ball 动量法及其最优个体收敛性

韦洪旭[1]　陇　盛[2]　陶　蔚[3]　陶　卿[1]
1 中国人民解放军陆军炮兵防空兵学院信息工程系　合肥 230031
2 国防科技大学系统工程学院大数据与决策实验室　长沙 410073
3 中国人民解放军军事科学院评估论证研究中心　北京 100091
(1140064271@qq.com)

**摘　要**　自适应策略与动量法是提升优化算法性能的常用方法。目前自适应梯度方法大多采用 AdaGrad 型策略,但该策略在约束优化中效果不佳,为此,研究人员提出了更适用于处理约束问题的 AdaGrad+方法,但其与 SGD 一样在非光滑凸情形下未达到最优个体收敛速率,结合 NAG 动量也并未达到预期的加速效果。针对上述问题,文中将 AdaGrad+调整步长的策略与 Heavy-Ball 型动量法加速收敛的优点相结合,提出了一种基于 AdaGrad+的自适应动量法。通过设置加权动量项、巧妙选取时变参数和灵活处理自适应矩阵,证明了该方法对于非光滑一般凸问题具有最优个体收敛速率。最后在$l_{\infty}$范数约束下,通过求解典型的 hinge 损失函数优化问题验证了理论分析的正确性,通过深度卷积神经网络训练实验验证了该方法在实际应用中也具有良好性能。

**关键词**:凸优化;自适应策略;AdaGrad+;Heavy-Ball 动量方法;收敛速率
**中图法分类号**　TP181

# Adaptive Heavy-Ball Momentum Method Based on AdaGrad+ and Its Optimal Individual Convergence

WEI Hongxu[1],LONG Sheng[2],TAO Wei[3] and TAO Qing[1]
1 Department of Information Engineering,Army Academy of Artillery and Air Defense of PLA,Hefei 230031,China
2 Laboratory for Big Data and Decision,College of Systems Engineering, National University of Defense Technology,Changsha 410073,China
3 Institute of Evaluation and Assessment Research,PLA Academy of Military Science,Beijing 100091,China

**Abstract**　Adaptive strategies and momentum methods are commonly used to improve the performance of optimization algorithms. Most of the adaptive gradient methods use the AdaGrad-type strategy at present. The AdaGrad+ method,which is more suitable for dealing with constrained problems,is proposed to solve the inefficiency of AdaGrad-type strategy on constrained optimization. But it is the same as SGD in non-smooth convex situations. The optimal individual convergence rate is not reached. Combining the strategy with NAG momentum but fail to achieve the expected acceleration effect. Aiming at the above problems,this paper proposes an adaptive momentum method based on AdaGrad+. The method uses the strategy of AdaGrad+ to adjust the step size,and inherits the advantages of the Heavy-Ball momentum method to accelerate the convergence. It is proved that the method achieves the optimal individual convergence rate for non-smooth convex problems by setting the weighted momentum term,selecting time-varying parameters skillfully and processing the adaptive matrix flexibly. Finally,experiments are conducted on the typical optimization problem of hinge loss function with $l_{\infty}$ norm constraint,and the experiment results verify the correctness of the theoretical analysis. In addition,the deep learning experiments confirm that the proposed method also has good performance in practical applications.

**Keywords**　Convex optimization,Adaptive strategy,AdaGrad+,Heavy-Ball momentum method,Convergence rate

机器学习问题普遍可以转化为求解目标函数最优解的优化问题。随机梯度下降法(SGD)[1]是解决优化问题的经典一阶优化算法,其计算代价低,且在实际应用中收敛速率快。目前,机器学习领域通常使用自适应步长策略和动量法两种优化策略来提升优化算法的性能。自适应步长策略指针对深度学习数据的稀疏性,根据数据不同维度的特征,利用迭代过程中的历史信息来调节每个维度参数的步长,降低对超参数的依赖性。动量方法即根据梯度的历史信息对解向量的更新方向进行调整,从而起到加速作用,提高算法在求解凸问题时的收敛速率,使其在非凸情形下可逃离鞍点和局部极值点[2-3]。将两种策略结合得到的自适应动量法不仅具有更快的收敛速率,而且弥补了 SGD 在步长选择

到稿日期:2022-09-14　返修日期:2023-07-05
通信作者:陶卿(qing.tao@ia.ac.cn)

以及参数更新方向上存在的不足。

AdaGrad 型自适应策略源于 Duchi 等[4]于 2011 年提出的 AdaGrad 算法，该策略在步长分母上添加对角矩阵，该矩阵通过累积历史梯度平方信息，使不同维度参数更新步长得以有效区分。AdaGrad 算法在一般凸情况下可获得和梯度下降法一样 $O(\sqrt{t})$ 的 regret 界（$t$ 为迭代步数）。目前常见的 RMSprop[5]，AdaDelta[6]等算法都是基于 AdaGrad 型策略发展而来的，它们在步长更新中加入了指数移动平均（EMA），避免过度累积历史信息。但对于约束最小化目标，梯度在最优时是非零的，上述自适应策略在迭代后期很难有效地找到合适步长进而得到最优解。2017 年 Levy[7]提出了一种适用于约束优化且达到最优收敛速率的自适应方法，但该方法需要全局最优位于约束域内的条件。Alina 等[8]于 2021 年提出了 AdaGrad＋算法，AdaGrad＋使用新的自适应步长矩阵，不直接使用历史梯度信息，而是利用前一步参数迭代信息更新自适应步长，使之在约束优化中具有良好表现。AdaGrad＋算法在光滑情况下可以得到 $O(1/t)$ 的收敛速率，但在非光滑情况下存在与 SGD 类似的问题，只能得到 $O(\sqrt{\ln t}/\sqrt{t})$ 的收敛速率，其相比非光滑凸情形的最优个体收敛速率存在 $\sqrt{\ln t}$ 因子的差距。

动量法在梯度下降法的基础上添加动量项，用于加快一阶梯度算法的收敛速度。目前常用的的动量方法有两类。一类是 Heavy-Ball 型动量法[9]，该方法由 Polyak 于 1964 年提出。Heavy-Ball 型动量法在目标函数各个方向变化强弱明显的情况下可以取得很好的加速效果，并且复杂度几乎没有增加。Ghadimi 等[10]对目标函数光滑条件下 Heavy-ball 方法的收敛性进行了深入的研究，得到了未达到最优的平均和个体收敛速率。Cheng 等[11]证明了在目标函数非光滑情况下 Heavy-Ball 型动量法具有 O1/$\sqrt{t}$ 的最优个体收敛速率。另一类是 Nesterov 提出的 NAG（Nesterov Accelerated Gradient）型动量法[3]。在目标函数光滑情况下，利用 NAG 型动量法的加速作用，能够得到 $O(1/t^2)$ 的收敛速率。在非光滑情况下，Tao 等[12-13]通过在投影次梯度中引进 NAG 调整步长的策略，得到 $O(1/\sqrt{t})$ 最优个体收敛速率。

为达到更好的收敛效果，将自适应策略与动量法相结合成为近年来的研究热门，称为自适应动量法。Kingma 等提出的 Adam 算法[14]是首个自适应动量法。Adam 在 SGD 基础上使用 AdaGrad 型自适应策略调整步长，并且用 EMA 型动量搜索方向，在实际应用中表现优异。虽然该算法证明了其在一般凸情形中具有 $O(1/\sqrt{t})$ 数据依赖的收敛速率，但 Reddi 等对此提出质疑，他们发现即使对于简单的凸优化问题，Adam 都无法收敛到全局极小值。为解决此问题，他们提出了 AMSGrad 和 AdamNC 两个修正算法[15]。为避免 Adam 存在的收敛性问题，Tao 等[16]放弃了对动量使用 EMA 策略，仅采用 EMA 策略调整自适应步长，提出了 Ada-HB 算法并证明了其具有 $O(1/\sqrt{t})$ 的最优个体收敛速率。Zhang 等[17]将 Ada-Belief 算法[18]调整步长的技巧和不采用 EMA 策略的 Heavy-Ball 型动量方法相结合，提出了 AdaBHB 算法，该算法在非光滑情况下同样具有最优个体收敛速率。而对于 AdaGrad＋策略的研究较少，文献[8]在 AdaGrad＋的基础上结合 NAG 动量法进行改进，在光滑情况下具有 $O(1/t^2)$ 的最优收敛速率，但在非光滑情况下只得到与 AdaGrad＋相同的 $O(\sqrt{\ln t}/\sqrt{t})$ 的收敛速率，未取得预想的加速效果。

上述分析表明采用 AdaGrad 策略的自适应动量法在约束优化中表现不佳，而现有的基于 AdaGrad＋的自适应 NAG 动量法未能达到最优个体收敛速率，因此，本文提出一种在满足在非光滑情形下适用于约束及无约束优化的方法。本文的主要工作如下：

（1）提出了一种 Ada-HB＋方法，将 AdaGrad＋自适应策略拓展到 Heavy-Ball 型动量法，弥补了以往自适应动量法在处理约束优化问题时的不足。

（2）证明了 Ada-HB＋算法在非光滑一般凸情形下具有最优个体收敛速率 $O(1/\sqrt{t})$（见定理 1），去除了 AdaGrad＋基础算法收敛速率中的 $\sqrt{\ln t}$ 因子，体现了 Heavy-Ball 的动量加速效果。

（3）通过约束条件下的非光滑凸优化实验验证了 Ada-HB＋收敛性分析的正确性，并将该方法应用于处理深度学习任务，证明其在实际运用中也具有良好性能。

值得指出的是，本文整体证明思路受文献[16]启发，但由于采用的自适应策略不同，在处理证明过程中涉及自适应矩阵的相关内容时区别很大。为得到非光滑个体收敛性，本文引入加权动量项，巧妙选取时变步长和动量项权重参数来转化迭代公式的形式（见引理 1）；使用 Zinkevich 处理在线优化问题收敛性的方法[19]来解决变步长与权重导致的递归问题；对于自适应矩阵的改变带来的证明上的困难，利用新矩阵正定的特征进行灵活处理（见引理 2）。

## 1 相关知识

本章主要对算法收敛速率、AdaGrad 算法、AdaGrad＋算法以及 Heavy-Ball 型动量法进行必要的介绍。

考虑如下优化问题：

$$\min_{\boldsymbol{w}\in Q} f(\boldsymbol{w}) \tag{1}$$

其中，$f(\boldsymbol{w})$ 为凸函数，$Q\subseteq R^n$ 为有界闭凸集，$\boldsymbol{w}^*$ 为式（1）的一个最优解。投影次梯度方法的迭代步骤为：

$$\boldsymbol{w}_{t+1}=\boldsymbol{P}_Q(\boldsymbol{w}_t-\alpha_t\boldsymbol{g}_t) \tag{2}$$

其中，$\boldsymbol{w}_t$ 为 $\boldsymbol{w}$ 在第 $t$ 步的迭代输出；$\alpha_t$ 为设置的衰减步长，一般凸情形中 $\alpha_t=\alpha/\sqrt{t}(\alpha>0)$；$\boldsymbol{g}_t$ 为 $f(\boldsymbol{w})$ 在 $\boldsymbol{w}_t$ 处的次梯度，$\boldsymbol{P}_Q$ 是在 $\boldsymbol{Q}$ 上的投影算子。

算法性能体现在空间复杂度和时间复杂度上，由于本文涉及的所有算法均为一阶梯度算法，因此具有相同的空间复杂度。而算法发展是在不增加每一步迭代的计算代价的基础上更加合理地利用了历史信息，因此不同算法的时间复杂度仅取决于达到特定精度所需的迭代步数，即收敛速率。收敛速率越快，算法复杂度越低。多数算法只给出以所有迭代平均方式作为输出时的平均收敛速率，然而人们更关注的是单步迭代作为输出时的个体收敛速率。对于投影次梯度等算法，平均收敛速率即 $f\bar{\boldsymbol{w}}_t-f(\boldsymbol{w}^*)$ 的收敛速率，其中 $\bar{\boldsymbol{w}}_t=1/t\sum_{k=1}^{t}\boldsymbol{w}_k$ 是算法历史迭代输出的算数平均值；个体收敛速率即

$f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)$的收敛速率，其中 $\boldsymbol{w}_t$ 为 $\boldsymbol{w}$ 在第 $t$ 步的迭代输出。对于非光滑优化问题，个体收敛相比平均收敛更难达到最优。Agarwal 等[20]证明了投影次梯度算法在非光滑凸情形下具有 $O1/\sqrt{t}$ 的最优平均收敛速率，而 Shamir 等[21]证明了 SGD 最优个体收敛速率只有 $O\ln t/\sqrt{t}$，这与非光滑凸问题的最优个体收敛速率还有对数阶的差距。

梯度下降法对于所有维度参数均使用同一步长，在稀疏优化中会影响算法的性能及收敛速率，因而自适应步长策略被提出。

AdaGrad 算法[4]的迭代公式如下：

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\alpha_t\boldsymbol{V}_t^{-1/2}\boldsymbol{g}_t \tag{3}$$

$$\boldsymbol{V}_t=\frac{1}{t}\sum_{i=1}^{t}\boldsymbol{g}_i^2 \tag{4}$$

其中 $\alpha_t=\alpha/\sqrt{t}$（$\alpha$ 为固定超参数），$\boldsymbol{g}_t$ 为次梯度，$\boldsymbol{g}_i^2=\operatorname{diag}\boldsymbol{g}_i\boldsymbol{g}_i^{\mathrm{T}}$ 为梯度外积矩阵的对角阵。

为了简单和直观起见，可以将 $\alpha_t\boldsymbol{V}_t^{-1/2}$ 整体视为步长。AdaGrad 算法即用带对角矩阵的自适应步长 $\alpha_t\boldsymbol{V}_t^{-1/2}$ 代替式(2)中的 $\alpha_t$。对角阵 $\boldsymbol{V}_t^{-1/2}$ 累积历史梯度平方信息，其对角线元素对应各维度参数的更新权重，使不同维度具有不同的步长。

AdaGrad+将 AdaGrad 推广到可行集 $\boldsymbol{Q}$ 是任意凸集的约束情况，在无约束情况下几乎与 AdaGrad 一致[8]，其迭代公式如下：

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\alpha_t\boldsymbol{D}_t^{-1}\boldsymbol{g}_t(\boldsymbol{w}\in\boldsymbol{Q}) \tag{5}$$

$$\boldsymbol{D}_{t+1,i}=\boldsymbol{D}_{t,i}\sqrt{1+\frac{\boldsymbol{w}_{t+1,i}-\boldsymbol{w}_{t,i}^2}{R_\infty^2}} \tag{6}$$

$$R_\infty\geqslant\max_{\boldsymbol{w},\boldsymbol{u}\in\boldsymbol{Q}}\|\boldsymbol{w}-\boldsymbol{u}\|_\infty \tag{7}$$

其中，$\boldsymbol{D}_t$ 为自适应对角矩阵且正定，根据式(6)更新对角线元素，$t$ 为迭代次数，$i$ 为元素维数，$R_\infty$ 为 $l_\infty$ 范数球的半径。

由式(4)和式(6)可以看出，AdaGrad+与 AdaGrad 的主要区别在于步长中的自适应矩阵不同，AdaGrad+不再直接累积历史梯度信息，而采用当前与上一步迭代参数差来更新权重。正如文献[8]所述，一方面，AdaGrad+利用前一步迭代信息可以更准确获取数据特征，从而找到合适的步长。当某一维度参数两次迭代的参数差值较大时，为避免因下降过快而错过最优值，该方向使用较小步长；当迭代差较小时，为提高收敛速率，则使用较大步长。另一方面，也可以避免过度使用历史信息导致迭代后期步长过小，出现停滞现象。

约束条件下，在达到域内最优值时梯度非零，通过梯度信息无法有效判断迭代是否接近最优值，因此 AdaGrad 很难调整到合适步长，找到最优解。然而当迭代收敛到最优值时，参数移动总是趋近于零。基于此，本文利用两步迭代参数差调整步长，使迭代参数差值不断趋近于零，即可有效找到最优解。

Heavy-Ball 型动量法有两种计算形式，一种以迭代前后两点的差为动量，另一种采用 EMA 形式计算动量。使用动量的 EMA 形式会导致收敛性分析中出现类似 Adam 无法收敛的问题[15]，并且在实际应用中还存在其他问题。使用 EMA 型动量在系数趋近于 1 时会导致算法的迭代陷入停滞，但传统 Heavy-Ball 动量不存在这样的问题[22]。故本文采用传统 Heavy-ball 型动量法，迭代步骤如下：

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\alpha_t\boldsymbol{g}_t+\beta_t(\boldsymbol{w}_t-\boldsymbol{w}_{t-1}) \tag{8}$$

其中，$\boldsymbol{w}_t-\boldsymbol{w}_{t-1}$ 为动量项，$\beta_t\in[0,1)$ 为动量系数。

## 2 Ada-HB+方法及其个体最优收敛性分析

本章提出 Ada-HB+算法，并给出约束条件下，目标函数非光滑一般凸情况下的个体收敛性证明。

### 2.1 Ada-HB+算法

将 AdaGrad+自适应策略与传统 Heavy-ball 型动量法相结合，提出更适用于约束优化的 Ada-HB+算法，考虑非光滑一般凸约束优化，$w\in Q$，其迭代形式为：

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\alpha_t\boldsymbol{D}_t^{-1}\boldsymbol{g}_t+\beta_t(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})$$

$$\boldsymbol{D}_{t+1,i}=\boldsymbol{D}_{t,i}\sqrt{1+\frac{\boldsymbol{w}_{t+1,i}-\boldsymbol{w}_{t,i}^2}{R_\infty^2}} \tag{9}$$

$$R_\infty\geqslant\max_{\boldsymbol{w},\boldsymbol{u}\in\boldsymbol{Q}}\|\boldsymbol{w}-\boldsymbol{u}\|_\infty$$

与以往大多自适应动量法不同的是，所提算法借鉴了 AdaGrad+自适应调整步长的策略，采用两步迭代的参数差来更新自适应矩阵，从而在处理约束优化问题时表现更佳。在此基础上添加不采用 EMA 形式的动量项 $\boldsymbol{w}_t-\boldsymbol{w}_{t-1}$，达到加速收敛的效果。

### 2.2 最优个体收敛性分析

本文的个体收敛性证明参考 Tao 等[16]对 Ada-HB 算法收敛性分析的思路，但 $\boldsymbol{V}_t$ 与 $\boldsymbol{D}_t$ 的不同导致它们在定理证明过程中对 $\sum_{k=1}^{t}\frac{1}{\sqrt{k}}\boldsymbol{g}_{k_{D_k^{-1}}}^2$ 的处理有所区别。

首先需要给出一个假设条件，该假设在以往收敛性分析中普遍存在。

假设 1(梯度有界)　存在一个常数 $M>0$ 使得：

$\boldsymbol{g}_{\infty t}\leqslant M(\forall\boldsymbol{w}\in\boldsymbol{Q})$

引进加权动量项 $\boldsymbol{p}_t=t(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})$，巧妙选取式(9)中的 $\alpha_t$ 和 $\beta_t$，将所提算法的迭代方式转化为引理 1 的形式，基于引理 1 可证明定理 1。

**引理 1**　令 $\boldsymbol{p}_t=t(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})$，$\alpha_t=\alpha/(t+2)\sqrt{t}$，$\beta_t=t/(t+2)$，则有：

$$\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1}=\boldsymbol{w}_t+\boldsymbol{p}_t-\frac{\alpha}{\sqrt{t}}\boldsymbol{D}_t^{-1}\boldsymbol{g}_t$$

证明：取 $\boldsymbol{p}_t=t(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})$ 得

$$\begin{aligned}\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1}&=\boldsymbol{w}_{t+1}+(t+1)(\boldsymbol{w}_{t+1}-\boldsymbol{w}_t)\\&=(t+2)\boldsymbol{w}_{t+1}-(t+1)\boldsymbol{w}_t\end{aligned}$$

将式(9)代入，得

$$\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1}=\boldsymbol{w}_t-(t+2)\alpha_t\boldsymbol{D}_t^{-1}\boldsymbol{g}_t+(t+2)\beta_t(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})$$

令 $\alpha_t=\alpha/(t+2)\sqrt{t}$，$\beta_t=t/(t+2)$，整理可得

$$\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1}=\boldsymbol{w}_t+\boldsymbol{p}_t-\frac{\alpha}{\sqrt{t}}D_t^{-1}\boldsymbol{g}_t$$

引理 1 得证。为解决定理 1 证明过程中变步长带来的递归问题，先证明引理 2。

**引理 2**　假设梯度有界，令 $R_\infty\geqslant\max_{\boldsymbol{w},\boldsymbol{u}\in\boldsymbol{Q}}\|\boldsymbol{w}-\boldsymbol{u}\|_\infty$，存在常数 $c$，使得 $D_{t,i}\geqslant c>0$（$D_{t,i}$ 为对角矩阵 $\boldsymbol{D}_t$ 的第 $i$ 维元素），则有：

$$\sum_{k=1}^{t}\sqrt{k}[(\boldsymbol{w}^*-(\boldsymbol{w}_k+\boldsymbol{p}_k)_{\boldsymbol{D}_k}^2-\boldsymbol{w}^*-(\boldsymbol{w}_{k+1}+p_{k+1})_{\boldsymbol{D}_k}^2)]+\sum_{k=1}^{t}$$

$$\frac{1}{\sqrt{k}}\boldsymbol{g}_{k_{D_k^{-1}}}^2 \leqslant R_\infty^2\sqrt{t}\sum_{i=1}^{d}D_{t,i}+\frac{M^2}{c}(2\sqrt{t}-1)$$

证明：借鉴 Zinkevich 处理在线优化时使用的迭代技巧进行整理。

$$\sum_{k=1}^{t}[\sqrt{k}(\|\boldsymbol{w}^*-(\boldsymbol{w}_k+\boldsymbol{p}_k)\|_{\boldsymbol{D}_k}^2)-\|\boldsymbol{w}^*-(\boldsymbol{w}_{k+1}+\boldsymbol{p}_{k+1})\|_{\boldsymbol{D}_k}^2]+\sum_{k=1}^{t}\frac{1}{\sqrt{k}}\|\boldsymbol{g}_k\|_{D_k^{-1}}^2$$

$$\leqslant\|\boldsymbol{w}^*-(\boldsymbol{w}_1+\boldsymbol{p}_1)\|_{\boldsymbol{D}_1}^2-\sqrt{t}\|\boldsymbol{w}^*-(\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1})\|_{\boldsymbol{D}_t}^2+\frac{M^2}{c}\sum_{k=1}^{t}\frac{1}{\sqrt{k}}+\sum_{k=2}^{t}[\sqrt{k}(\|\boldsymbol{w}^*-(\boldsymbol{w}_k+\boldsymbol{p}_k)\|_{\boldsymbol{D}_k}^2)-\sqrt{k-1}(\|\boldsymbol{w}^*-(\boldsymbol{w}_k+\boldsymbol{p}_k)\|_{\boldsymbol{D}_{k-1}}^2)]$$

$$\leqslant\sum_{i=1}^{d}\boldsymbol{D}_{1,i}(\boldsymbol{w}_i^*-(\boldsymbol{w}_{1,i}+\boldsymbol{p}_{1,i}))^2+\frac{M^2}{c}(2\sqrt{t}-1)+\sum_{i=1}^{d}\sum_{k=2}^{t}((\sqrt{k}D_{k,i}-\sqrt{k-1}\boldsymbol{D}_{k-1,i})(\boldsymbol{w}_i^*-(\boldsymbol{w}_{k,i}+\boldsymbol{p}_{k,i}))^2)$$

$$\leqslant R_\infty^2\sum_{i=1}^{d}\boldsymbol{D}_{1,i}+\sum_{i=1}^{d}\sum_{k=2}^{t}[R_\infty^2(\sqrt{k}\boldsymbol{D}_{k,i}-\sqrt{k-1}\boldsymbol{D}_{k-1,i})]+\frac{M^2}{c}(2\sqrt{t}-1)=R_\infty^2\sqrt{t}\sum_{i=1}^{d}\boldsymbol{D}_{t,i}+\frac{M^2}{c}(2\sqrt{t}-1)$$

引理 2 得证。

**定理 1** 设 $f(w)$为一般凸函数，取 $\alpha_t=\alpha/(t+2)\sqrt{t}$，$\beta_t=t/(t+2)$，$w_t$ 由式(9)生成，存在常数 $c$，使得 $D_{t,i}\geqslant c>0$，则：

$$f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)\leqslant\frac{f(\boldsymbol{w}_0)-f(\boldsymbol{w}^*)}{1+t}+\frac{R_\infty^2\sqrt{t}}{2\alpha(t+1)}\sum_{i=1}^{d}D_{t,i}+\frac{\alpha M^2}{2c1+t}(2\sqrt{t}-1)$$

证明：由引理 1 得

$\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1}=\boldsymbol{w}_t+\boldsymbol{p}_t-\frac{\alpha}{\sqrt{t}}\boldsymbol{D}_t^{-1}\boldsymbol{g}_t$ 从而有

$$\boldsymbol{w}^*-(\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1})_{\boldsymbol{D}_t}^2=\boldsymbol{w}^*-(\boldsymbol{w}_t+\boldsymbol{p}_t)+\frac{\alpha}{\sqrt{t}}\boldsymbol{D}_t^{-1}\boldsymbol{g}_{t\,\boldsymbol{D}_t}^2$$

$$=\boldsymbol{w}^*-(\boldsymbol{w}_t+\boldsymbol{p}_t)_{\boldsymbol{D}_t}^2+\frac{\alpha}{\sqrt{t}}\boldsymbol{g}_{t\,D_t^{-1}}^2+2\boldsymbol{w}^*-(\boldsymbol{w}_t+\boldsymbol{p}_t),\frac{\alpha}{\sqrt{t}}\boldsymbol{g}_t$$

将 $\boldsymbol{p}_t=t(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})$代入上式第三项，得

$$\boldsymbol{w}^*-(\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1})_{\boldsymbol{D}_t}^2\leqslant\boldsymbol{w}^*-(\boldsymbol{w}_t+\boldsymbol{p}_t)_{\boldsymbol{D}_t}^2+\frac{\alpha}{\sqrt{t}}\boldsymbol{g}_{t_{D_t^{-1}}}^2+2\boldsymbol{w}^*-\boldsymbol{w}_t,\frac{\alpha}{\sqrt{t}}\boldsymbol{g}_t+2\boldsymbol{w}_{t-1}-\boldsymbol{w}_t,\frac{\alpha}{\sqrt{t}}\boldsymbol{g}_t$$

$f(\boldsymbol{w})$为一般凸函数，由凸函数的一阶判别条件得

$\boldsymbol{w}^*-\boldsymbol{w}_t,g_t\leqslant f(\boldsymbol{w}^*)-f(\boldsymbol{w}_t),\boldsymbol{w}_{t-1}-\boldsymbol{w}_t$，

$\boldsymbol{g}_t\leqslant f(\boldsymbol{w}_{t-1})-f(\boldsymbol{w}_t)$

代入上式并将不等式两边同乘$\sqrt{t}/2\alpha$，得

$$\frac{\sqrt{t}}{2\alpha}\boldsymbol{w}^*-(\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1})_{\boldsymbol{D}_t}^2\leqslant\frac{\sqrt{t}}{2\alpha}\boldsymbol{w}^*-(\boldsymbol{w}_t+\boldsymbol{p}_t)_{\boldsymbol{D}_t}^2+\frac{\alpha}{2\sqrt{t}}g_{t_{D_t^{-1}}}^2+f(\boldsymbol{w}^*)-f(\boldsymbol{w}_t)+t(f(\boldsymbol{w}_{t-1})-f(\boldsymbol{w}_t))$$

将上式从 $k=1$ 到 $t$ 累加，得

$$(t+1)[f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)]\leqslant f(\boldsymbol{w}_0)-f(\boldsymbol{w}^*)+\sum_{k=1}^{t}\frac{\alpha}{2\sqrt{k}}\boldsymbol{g}_{k_{D_t^{-1}}}^2+\sum_{k=1}^{t}[\frac{\sqrt{k}}{2\alpha}(\boldsymbol{w}^*-(\boldsymbol{w}_k+\boldsymbol{p}_k)_{\boldsymbol{D}_k}^2-\boldsymbol{w}^*-(\boldsymbol{w}_{k+1}+\boldsymbol{p}_{k+1})_{\boldsymbol{D}_k}^2)]$$

根据引理 2，可得

$$f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)\leqslant\frac{f(\boldsymbol{w}_0)-f(\boldsymbol{w}^*)}{1+t}+\frac{R_\infty^2\sqrt{t}}{2\alpha1+t}\sum_{i=1}^{d}\boldsymbol{D}_{t,i}+\frac{\alpha M^2}{2c1+t}(2\sqrt{t}-1)$$

定理 1 得证。

**推论 1** 设 $f(\boldsymbol{w})$为一般凸函数，取 $\alpha_t=\alpha/(t+2)\sqrt{t}$，$\beta_t=t/(t+2)$，$\boldsymbol{w}_t$ 由式(9)生成，存在常数 $c$，使得 $D_{t,i}\geqslant c>0$，则：

$$f\left(\frac{1}{t}\sum_{k=1}^{t}\boldsymbol{w}_t\right)-f(\boldsymbol{w}^*)\leqslant O\left(\frac{1}{\sqrt{t}}\right)$$

由推论 1 可以看出平均收敛速率相较于个体收敛速率更易获得。综上可知，在非光滑一般凸的条件下 Ada-HB+算法可以得到 $O(1/\sqrt{t})$的最优个体收敛速率。然而上述证明是在批处理条件下完成的，批处理并不适用于大规模数据集。为解决此问题，我们将上述算法推广至随机形式。

假设一个二分类问题的训练样本集为 $S=\{(\boldsymbol{x}_i,y_i)\mid i=1,2,\cdots,m\}\subseteq R^n\times\{+1,1\}$，其中$(\boldsymbol{x}_i,y_i)$满足独立同分布。采用“hinge 损失”作为非光滑优化问题的损失函数，即：$f_i(\boldsymbol{w})=\max\{0,1-y_i\boldsymbol{w},\boldsymbol{x}_i\}$，则优化目标函数为：

$$\min_{w\in Q}f(\boldsymbol{w})=\frac{1}{m}\sum_{i=1}^{m}f_i(\boldsymbol{w})\tag{10}$$

约束情况下随机形式 Ada-HB+算法的迭代步骤则可以表示为：

$$\boldsymbol{w}_{t+1}=P_Q[\boldsymbol{w}_t-\alpha_t\boldsymbol{D}_t^{-1}\nabla f_i(\boldsymbol{w}_t)+\beta_t(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})]\tag{11}$$

本文中采用文献[23]的方式计算 hinge 损失函数次梯度：

$$\nabla f_i(\boldsymbol{w}_t)=\frac{1}{m}\sum_{(\boldsymbol{x}_i,y_i)\in A_t^+}y_i\boldsymbol{x}_i\tag{12}$$

其中，$A_t\subseteq S$，$A_t^+=\{(\boldsymbol{x}_i,y_i)\in A_t:y_i\boldsymbol{w},\boldsymbol{x}_i<1\}$，在实验中设置$|A_t|=1$。

随机形式的 Ada-HB+算法具体执行步骤如算法 1 所示。

**算法 1** AdaHB+算法

输入：循环次数 t

输出：$\mathbf{w}_t$

1. 初始化 向量 $\mathbf{w}_1\in Q$，$\mathbf{D}_1=\mathbf{I}$

2. For k=1 to t

3. 可能选取 i∈{1,2,3,…,m}

4. 根据式(12)计算次梯度$\nabla f_i(\mathbf{w}_k)$

5. 取 $\alpha_k=\alpha/(k+2)\sqrt{k}$，$\beta_k=k/(k+2)$

6. 由式(11)计算 $\mathbf{w}_{k+1}$

当样本满足独立同分布的条件时，通过随机抽取方式计算得到的$\nabla f_i(\boldsymbol{w}_t)$即为函数 $f(\boldsymbol{w})$在 $\boldsymbol{w}_t$ 处次梯度的无偏估计。由上述算法可以看出，随机形式就是用目标函数梯度的无偏估计替换批处理中的梯度。Rakhlin 等[24]给出了将收敛界从批处理形式转换到随机形式的技巧。与文献[11]类似，可以将本文的定理 1 转换为随机形式的定理 2。

**定理 2** 设 $f(\boldsymbol{w})$为一般凸函数，取 $\alpha_t=\alpha/(t+2)\sqrt{t}$，$\beta_t=t/(t+2)$，$\boldsymbol{w}_t$ 由式(11)生成，存在常数 $c$，使得 $D_{t,i}\geqslant c>0$，则：

$$E([f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)])\leqslant\frac{f(\boldsymbol{w}_0)-f(\boldsymbol{w}^*)}{1+t}+\frac{R_\infty^2\sqrt{t}}{2\alpha(t+1)}\sum_{i=1}^{d}D_{t,i}+\frac{\alpha M^2}{2c1+t}(2\sqrt{t}-1)$$

由定理 2 可知，在非光滑条件下随机 AdaHB+算法可以达到 $O(1/\sqrt{t})$的最优个体收敛速率。

## 3 实验及结果分析

AdaGrad+算法在约束优化中的实际性能已经在文献[8]中得到了充分的实验验证，本文主要证明 Ada-HB+算法具有最优收敛速率。因此本章重点对 Ada-HB+算法的收敛性进行验证，通过约束凸优化实验验证 Ada-HB+算法个体收敛速率理论分析的正确性，并通过深度学习实验验证该算法在实际任务中具有良好的性能。

两组实验均采用 6 种优化算法进行比较，为更好地体现本文的研究主旨，我们针对性地选择几种算法，分别为 Heavy-Ball 算法、AdaGrad 算法，SGD 算法、Adam 算法、基础 AdaGrad+ 算法以及 Ada-HB+算法。上述算法都是某一类方法的典型代表，对比意义较大。共同超参数 $\alpha$ 从 $\{1.5,1,0.8,0.1,0.01,0.001\}$的集合中选取，其他参数采用以往文献中该算法实验表现较好的参数设置，具体参数设置如表 1 所列。

表 1 参数设置

Table 1 Parameter settings

| 算法 | 参数 |
|---|---|
| SGD | $\alpha_t=\alpha/\sqrt{t}$ |
| Heavy-Ball | $\alpha_t=\alpha/(t+2)\sqrt{t},\beta_t=t/(t+2)$ |
| AdaGrad | $\alpha_t=\alpha/\sqrt{t},\varepsilon=1\times10^{-8}$ |
| Adam | $\alpha_t=\alpha/\sqrt{t},\varepsilon=1\times10^{-8},\beta_1=0.9,\beta_2=0.999$ |
| AdaGrad+ | $\alpha_t=\alpha/\sqrt{t}$ |
| Ada-HB+ | $\alpha_t=\alpha/(t+2)\sqrt{t},\beta_t=t/(t+2)$ |

优化算法的退出条件大多需要计算目标函数值，而计算目标函数需要遍历所有样本，代价较大。对于求解大规模优化问题，机器学习优化方法多采用随机形式，只使用部分样本，无须计算目标函数值，因此通常没有退出条件，都是以迭代次数为标准，故本文在实验中仅根据以往经验设置合适的迭代次数。为了对比更加公平，每个算法在 6 个数据集上均运行 5 次，取平均值作为最后的输出。

### 3.1 一般凸函数分类优化实验

该实验通过处理约束条件下 hinge 损失分类优化问题验证了个体收敛性分析的正确性，约束区域为 $l_\infty$ 范数球$\{w_{l_\infty}\leqslant z\}$。使用 SLEP 工具箱[25]中的 eplb 函数来实现 $l_\infty$ 范数球的投影操作。由于数据集不同，$z$ 也应选取不同的值，但在同一数据集中，各算法选取相同的 $z$。实验采用了 astro，CCAT，w8a，ijcnn1，rcv1，covtype 这 6 个常见标准数据集，这些数据集均来自于 LIBSVM 网站，详细数据如表 2 所列。

表 2 标准数据集介绍

Table 2 Introductionto standard datasets

| 数据集 | 训练样本集 | 维数 |
|---|---|---|
| astro | 29882 | 99757 |
| CCAT | 23149 | 47236 |
| w8a | 49749 | 300 |
| ijcnn1 | 49990 | 22 |
| rcv1 | 20242 | 47236 |
| covtype | 522911 | 54 |

图 1 为在 6 个标准数据集下 6 种算法的收敛速率对比图，图中横坐标表示迭代步数，纵坐标为当前目标函数值与最优目标函数值的差。由图可见，在约束凸优化情况下，迭代 10000 步后，6 种算法在 6 个标准数据集上的相对目标函数值都达到了 $10^{-4}$的精度，且具有基本相同的收敛趋势。而在同一精度下，Ada-HB+收敛速度相对较快，在迭代步数相同时，Ada-HB+相对目标函数值最低，与理论分析相吻合。

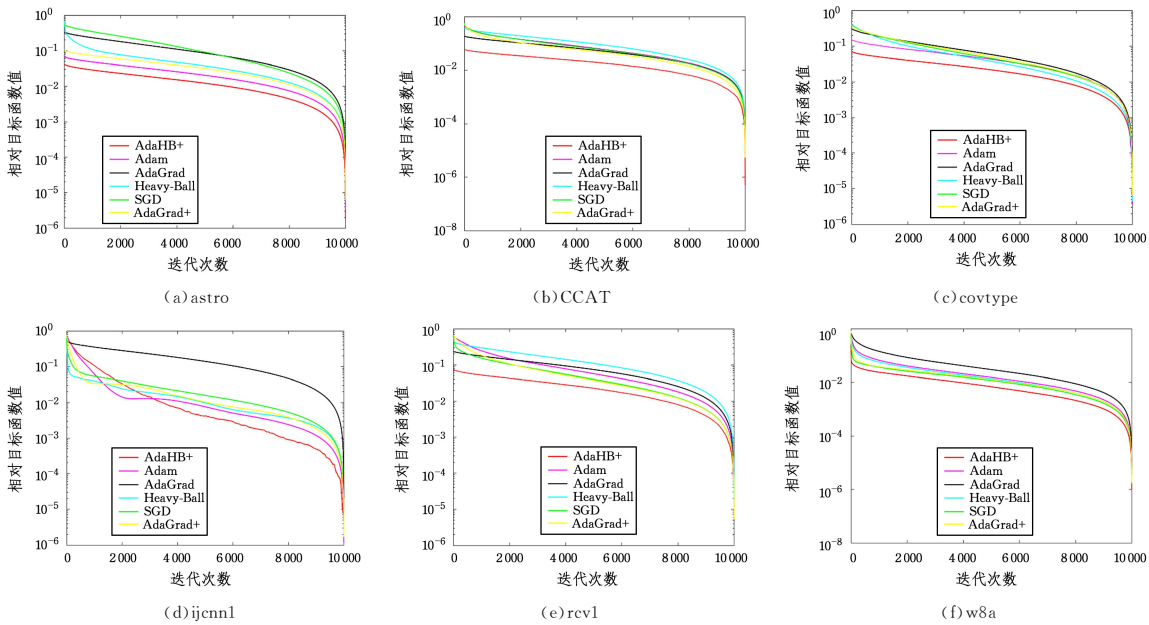


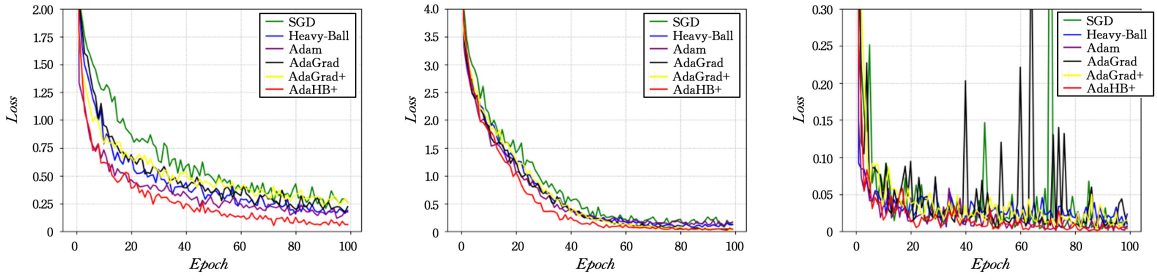

图 1 收敛速率对比

Fig. 1 Comparison of convergence rates

### 3.2　深度卷积神经网络训练实验

本组实验的目的是验证 Ada-HB＋算法在实际应用中的性能。按照 Tao 等[16]的思路，在典型残差网络 Res-Net18 上进行实验，实验中采用参数权重衰减和批量归一化策略来减少过拟合，所用的损失函数为交叉熵损失。

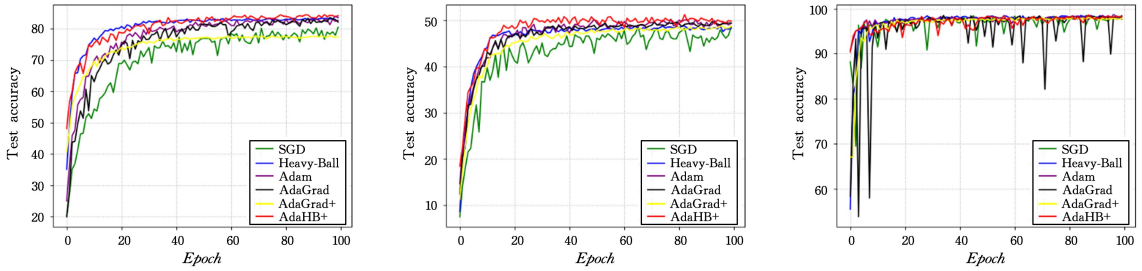
使用 6 种算法分别在 3 个公开的标准数据集上进行对比实验，数据集分别为 CIFAR10，CIFAR100，MNIST。图 2 为各算法在 3 种数据集上的损失对比图，图 3 为各算法在 3 种数据集上的测试精度对比图，其中横坐标为迭代次数，纵坐标分别为损失值和测试精度。可以看出，在各数据集上，相同迭代步数下，Ada-HB＋的训练损失相对较低，在测试精度上也略有优势。在其他深度学习网络中进行的验证实验中，Ada-HB＋同样取得了较优的实验效果。由于篇幅限制，文中仅展示 Res-Net18 上的结果。由此可见，本文所提自适应 HB 动量法在处理实际深度学习任务时具有良好的性能。




(a)CIFAR10　(b)CIFAR100　(c)MNIST

图 2　Res-Net18 损失对比

Fig. 2　Comparison of loss values on Res-Net18




(a)CIFAR10　(b)CIFAR100　(c)MNIST

图 3　Res-Net18 测试精度对比

Fig. 3　Comparison of test accuracies on Res-Net18

**结束语**　本文结合 AdaGrad＋自适应策略以及传统 Heavy-Ball 动量，提出了一种名为 Ada-HB＋的自适应动量方法，通过理论分析证明了 Ada-HB＋在非光滑一般凸条件下具有 $O1/\sqrt{t}$ 的最优个体收敛速率，体现了动量的加速特点。与基于 AdaGrad 策略的自适应动量法相比，Ada-HB＋继承了 AdaGrad＋适用于处理约束问题的优点。实验验证了理论分析的正确性以及所提方法在实际应用中的良好性能。然而，本文只分析了非光滑一般凸情况，并未涉及强凸情形，下一步将继续对强凸条件下 Ada-HB＋算法的收敛性以及在深度学习中的应用进行研究。

## 参 考 文 献

[1] ROBBINS H，MONRO S. A Stochastic Approximation Method [J]. The Annals of Mathematical Statistics，1951，22(3)：400-407.

[2] SUN T，LI D S，QUAN Z，et al. Heavy-ball Algorithms Always Escape Saddle Points[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI 2019). San Francisco：Morgan Kaufmann，2019：3520-3526.

[3] NESTEROV Y. A Method of Solving a Convex Programming Problem with Convergence Rate $O(1/k^2)$[J]. Soviet Mathematics Doklady，1983，27(2)：372-376.

[4] DUCHI J，HAZAN E，SINGER Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization[J]. Journal of Machine Learning Research，2011，12：2121-2159.

[5] TIELEMAN T，HINTON G. RMSProp：Divide the Gradient by a Running Average of its Recent Magnitude[D]. Toronto：University of Toronto，2012.

[6] MATTHEW D. ZEILE R. ADADELTA：An Adaptive Learning Rate Method[J]. arXiv：1212. 5701，2012.

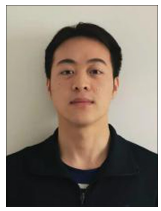
[7] LEVY K Y. Online to offline conversions，universality and adaptive minibatch sizes[C]// Proceedings of the 31st Annual Conf. on Neural Information Processing Systems (NeurIPS). Cambridge，MA：MIT Press，2017：1613-1622.

[8] ALINE E，HUY L，ADRIAN V. Adaptive Gradient Methods for Constrained Convex Optimization and Variational Inequalities [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021：7314-7321.

[9] POLYAK B T. Some Methods of Speeding up the Convergence of Iteration Methods[J]. USSR Computational Mathematics and Mathematical Physics，1964，4(5)：1-17.

[10] GHADIMI E，FEYZMAHDAVIAN H R，JOHANSSON M. Global Convergence of the Heavy-Ball Method for Convex Optimization

[C]// Proceedings of the European Control Conference. Washington, USA: IEEE, 2015: 310-315.
[11] CHENG Y J, TAO W, LIU Y X, et al. Optimal Iindividual Convergence Rate of the Heavy-Ball-based Momentum Methods [J]. Chinese Journal of Computer Research and Development, 2019, 56(8): 1686-1694.
[12] TAO W, PAN Z S, CHU D J, et al. The Individual Convergence of Projected Subgradient Methods Using the Nesterov's Step-Size Strategy[J]. Chinese Journal of Computers, 2018, 41(1): 164-176.
[13] TAO W, PAN Z S, WU G W, et al. The Strength of Nesterov's Extrapolation in the Individual Convergence of Non smooth Optimization[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(7): 2557-2568.
[14] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization[J]. arXiv: 1412. 6980. 2014.
[15] REDDI S J, KALE S, KUMAR S. On the Convergence of Adam and Beyond[J]. arXiv: 1904. 09237, 2019.
[16] TAO W, LONG S, WU G W, et al. The Role of Momentum Parameters in the Optimal Convergence of Adaptive Polyak's Heavy-Ball Methods[J]. arXiv: 2102. 07314, 2021.
[17] ZHANG Z D, LONG S, BAO L, et al. AdaBelief Based Heavy-Ball Momentum Method[J]. Pattern Recognition and Artificial Intelligence, 2022, 35(2): 106-115.
[18] ZHUANG J T, TANG T, DING Y F, et al. AdaBelief Optimizer: Adapting Stepsizes by the Belief in Observed Gradients [J]. arXiv: 2010. 07468, 2020.
[19] ZINKEVICH M. Online Convex Programming and Generalized Infinitesimal Gradient Ascent[C]// Proceedings of the 20th International Conference on Machine Learning. New York: ACM, 2003.
[20] AGARWAL A, BARTLETT P L, RAVIKUMAR P, et al. Information-Theoretic Lower Bounds on the Oracle Complexity of Stochastic Convex Optimization[J]. IEEE Transactions on Information Theory, 2012, 58(5): 3235-3249.
[21] SHAMIR O, ZHANG T. Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes[C]// Proceedings of the 29th International Conference on Machine Learning. New York, USA: ACM, 2013: 71-79.
[22] ZOU F Y, SHEN L, JIE Z Q, et al. A Sufficient Condition for Convergences of Adam and RMSprop[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA. 2019: 11119-11127.
[23] SHALEV S S, SINGER Y, SREBRO N, et al. Pegasos: Primal estimated sub-gradient solver for svm[J]. Mathematical Programming, 2011, 127(1): 3-30.
[24] RAKHLIN A, SHAMIR O, SRIDHARAN K. Making gradient descent optimal for strongly convex stochastic optimization [C]// Proceedings s of the 29th International Conference on Machine Learning. New York, USA: ACM, 2012: 449-456.
[25] LIU J, JI S, YE J. SLEP: Sparse learning with efficient projections[D]. Arizona: Arizona State University, 2009.



**WEI Hongxu**, born in 1998, postgraduate. His main research interests include pattern recognition and machine learning.

**TAO Qing**, born in 1965, Ph. D, professor, Ph. D supervisor, is a member of China Computer Federation. His main research interests include pattern recognition, machine learning and applied mathematics.

（责任编辑：何杨）