软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*, 2022,33(4):1231−1243 [doi: 10.13328/j.cnki.jos.006464]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 基于 AdaGrad 的自适应 NAG 方法及其最优个体收敛性*

陇 盛[1], 陶 蔚[2], 张泽东[1], 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, 安徽 合肥 230031)

[2](中国人民解放军军事科学院 战略评估咨询中心, 北京 100091)

通信作者: 陶卿, E-mail: qing.tao@ia.ac.cn

**摘 要**: 与梯度下降法相比, 自适应梯度下降方法(AdaGrad)利用过往平方梯度的算数平均保存了历史数据的几何信息, 在处理稀疏数据时获得了更紧的收敛界. 另一方面, Nesterov 加速梯度方法(Nesterov's accelerated gradient, NAG)在梯度下降法的基础上添加了动量运算, 在求解光滑凸优化问题时具有数量级加速收敛的性能, 在处理非光滑凸问题时也获得了最优的个体收敛速率. 最近, 已经出现了自适应策略与 NAG 相结合的研究, 但现有代表性的自适应 NAG 方法 AcceleGrad 由于采取的自适应方式与 AdaGrad 不同, 步长未能在不同维度上体现差异性, 仅得到了加权平均方式的收敛速率, 个体收敛速率的理论分析尚存在缺失. 提出了一种自适应 NAG 方法, 继承了 AdaGrad 的步长设置方式, 证明了所提算法在解决约束非光滑凸优化问题时具有最优的个体收敛速率. 在 $L1$ 范数约束下, 通过求解典型的 hinge 损失函数分类和 $L1$ 损失函数回归优化问题. 实验验证了理论分析的正确性, 也表明了所提算法的性能优于 AcceleGrad.

**关键词**: 机器学习; 凸优化; 自适应算法; NAG 方法; 个体收敛速率

**中图法分类号**: TP18

中文引用格式: 陇盛, 陶蔚, 张泽东, 陶卿. 基于 AdaGrad 的自适应 NAG 方法及其最优个体收敛性. 软件学报, 2022, 33(4): 1231−1243. http://www.jos.org.cn/1000-9825/6464.htm

英文引用格式: Long S, Tao W, Zhang ZD, Tao Q. Adaptive NAG Methods Based on AdaGrad and Its Optimal Individual Convergence. Ruan Jian Xue Bao/Journal of Software, 2022, 33(4): 1231−1243 (in Chinese). http://www.jos.org.cn/1000-9825/6464.htm

## Adaptive NAG Methods Based on AdaGrad and Its Optimal Individual Convergence

LONG Sheng[1], TAO Wei[2], ZHANG Ze-Dong[1], TAO Qing[1]

[1](Department of Information Engineering, PLA Army Academy of Artillery and Air Defense, Hefei 230031, China)

[2](Center for Strategic Assessment and Consulting, PLA Academy of Military Sciences, Beijing 100091, China)

**Abstract**: Compared with the gradient descent, the adaptive gradient descent (AdaGrad) keeps the geometric information of historical data by using the arithmetic average of squared past gradients, and obtains tighter convergence bounds when copingwith sparse data. On the other hand, by adding the momentum term to gradient descent, Nesterov's accelerated gradient (NAG) not only obtains order of magnitude accelerated convergence for solving smooth convex optimization problems, but also achieves the optimal individual convergence rate for non-smooth convex problems. Recently, there have been studies on the combination of adaptive strategy and NAG. However, as a typical adaptive NAG algorithm, AcceleGrad fails to reflect the distinctions between dimensions due to using different adaptive variant from AdaGrad, and it only obtains the weighted averaging convergence rate. So far, there still lacks the theoretical analysis of individual convergence rate. In this study, an adaptive NAG method, which inherits AdaGrad's step size setting, is proposed. It

* 基金项目: 国家自然科学基金(62076252, 62106281)
本文由"面向开放场景的鲁棒机器学习"专刊特约编辑陈恩红教授、李宇峰副教授、邹权教授推荐.
收稿时间: 2021-03-09; 修改时间: 2021-07-16; 采用时间: 2021-08-27; jos 在线出版时间: 2021-10-26

is proved that the proposed algorithm attains the optimal individual convergence rate when solving the constrained non-smooth convex optimization problems. The experiments are conducted on the typical optimization problem of hinge loss function for classification and $L1$ loss function for regression with $L1$ norm constraint, and experimental results verify the correctness of the theoretical analysis and superior performance of the proposed algorithm over AcceleGrad.

**Key words**: machine learning; convex optimization; adaptive algorithm; NAG method; individual convergence rate

机器学习模型虽然有多种形式, 但最终都需要进行模型参数的优化求解并判断算法在参数寻优过程中是否收敛以及分析其收敛速率. 实际上, 优化算法的收敛性分析一直是机器学习不可或缺的关键环节. 机器学习领域早期的做法是: 先用 Regret bound 描述在线算法的收敛界[1], 然后根据 online-to-batch 转换技巧得到随机情形下的收敛速率[2]. 上述分析方式只能得到算法以所有迭代平均方式作为输出时的收敛速率(即平均收敛速率). 然而, 人们往往关注的是单步迭代作为输出时的收敛速率(个体收敛速率), 并且个体解更利于保持机器学习正则化项的结构. 对于凸优化问题, 从个体解的最优收敛性能够方便获得最优平均收敛性, 但反之却不成立. 2012 年, Shamir 等人从非光滑凸优化 SGD(stochastic gradient descent)的最优平均收敛速率 $O(1/\sqrt{t})$ 出发, 仅仅得到了 $O(\log t/\sqrt{t})$ 的个体收敛速率[3]; 而且近期的研究也表明, SGD 的最优个体收敛速率就是 $O(\log t/\sqrt{t})$[4−6], 其中, $t$ 是迭代步数. 显然, 能否改进 SGD, 使其获得最优的个体收敛性值得探索.

自适应调整步长和添加动量运算调整方向, 是进一步提升 SGD 性能的两种主要技巧. 基于处理稀疏数据零分量时可以采用更大步长这一直观的思路, 2011 年, Duchi 等人提出了 AdaGrad 算法[7], 它可以看作是 SGD 与自适应步长策略的首次结合. 在处理一般凸优化问题时, 为了获得最优的平均收敛性, SGD 通常取步长 $O(1/\sqrt{t})$. 与 SGD 所不同的是, AdaGrad 在步长的分母上添加了一个对角矩阵. 该矩阵以算数平均方式积累过往平方梯度的信息, 不同维度参数间每次迭代的步长因矩阵对角线上元素的不同而得到了有效的区分. Duchi 等人证明了在线 AdaGrad 在处理一般凸问题时具有 $O(\sqrt{t})$ Regret bound, 达到了和 SGD 一样的最优收敛速率; 但在处理稀疏数据时, AdaGrad 算法却能获得比 SGD 更小因子的收敛界. AdaGrad 通过用对角矩阵记录历史数据, 缓和了算法对超参数的过度依赖性, 能够满足不同维度参数对不同步长的需求. 随后, 受深度学习优化算法发展的驱动, 涌现出多种形式的自适应算法, 如 AdaDelta[8], Adam[9]等.

基于动量的优化方法主要有两种: 一种是 Polyak 于 1964 年提出的 Heavy-ball 方法[10], 另一种是 Nesterov 在 1983 年提出的加速梯度(Nesterov's accelerated gradient, NAG)方法[11]. NAG 早期的研究大多集中于光滑凸优化问题, 它填补了"一阶梯度下降法在当时处理光滑凸函数只有 $O(1/t)$收敛速率"与 Nemirovski 和 Yudin 所证明的"任何一阶优化算法都不可能得到比 $O(1/t^2)$更快的收敛速率"之间的间隙[12]. Tseng 在文献[13]中给出了 NAG 方法特殊情况下等价的 3 种形式, 在统一的框架下, 证明了 3 种形式 NAG 的最优收敛性. 对于非光滑凸优化问题, Lan 等人在假设总迭代次数固定的前提下, 对第 2 种形式 NAG 方法得到了最优个体收敛速率, 但无法应用于大规模数据的增量学习[14]; Devolder 等人得到了第 3 种形式 NAG 方法的最优个体收敛速率, 但其需要两步的梯度运算, 缺失了一阶梯度算法原有的直观可解释性, 并且也不适用于包含复杂梯度运算的优化问题[15]. 最近, 文献[16]将第 1 种形式 NAG 方法推广到非光滑问题中, 对一般凸和强凸问题均得到了最优的个体收敛速率. 在此基础上, 文献[17]进一步考虑了 NAG 方法在处理有偏差梯度时的个体收敛速率.

最近, 出现了很多将自适应步长策略和 NAG 动量两种技巧相结合以获取更好收敛性能的研究工作. 2015 年, Dozat 将 Adam 的 Heavy-ball 型动量替换成 NAG 型动量, 提出了 NAdam[18]. 虽然 NAdam 取得了很好的实验效果, 但却没有进行必要的收敛性分析. 随后, 为了能不做任何修改使算法在光滑、非光滑和随机情况下同时保证收敛性, Levy 等人针对无约束问题提出了 AcceleGrad 算法[19]. AcceleGrad 由文献[20]采取的自适应步长策略结合文献[21]中 NAG 方法, 并且以所有迭代的加权平均作为算法输出得到. 虽然不需要事先知道光滑先验信息就能在光滑条件下达到最优收敛速率 $O(1/t^2)$, 但是对非光滑问题只得到了次优加权平均收敛速率 $O(\sqrt{\log t}/\sqrt{t})$. 为解决 AcceleGrad 不适用于约束条件的 open 问题以及去掉其非光滑收敛界上的对数因子, Kavis 等人继承了 AcceleGrad 的自适应步长策略, 用 Mirror-Prox 方法替换 NAG 动量, 得到了 UniXGrad[22]算

法, 其在光滑和非光滑情况下都分别得到了最优收敛速率. 但是算法依然是以加权平均形式输出的, 因此不具有个体解的优点. 需要指出的是: AcceleGrad 和 UniXGrad 的自适应步长策略与 AdaGrad 的风格是不同的, 这种区别主要体现在 AcceleGrad 和 UniXGrad 没有使用矩阵, 而只是使用梯度累加和与其他项组合作为步长. 这样复杂的设计虽然能构造普适算法应对多种形式的问题, 并且步长也符合常规的 $O(1/\sqrt{t})$, 但是失去了 AdaGrad 自适应步长策略能体现不同维度差异的特性, 因此该算法可能不是处理高维数据的最佳选择. 幸运的是, 文献[23]提出了 NAG 动量与多种自适应步长策略(包括 AdaGrad 等)结合的统一框架 A2GRAD, 参数取值不同对应累积历史梯度方式不同, 并且所得系列算法在光滑条件下都能达到 $O(1/t^2)$最优收敛速率, 遗憾的是没有进一步讨论非光滑情况.

值得指出的是: 最近的文献[24]剖析了 Heavy-ball 型动量参数在优化算法中扮演的角色, 将指数移动平均(EMA)型自适应策略与 Heavy-ball 动量方法进行了有效的结合. AdaGrad 步长策略作为 EMA 特殊情况得到继承, 同时又保证了 Heavy-ball 动量方法的最优个体收敛性[25]. 受这项工作的启发, 又由于 AcceleGrad 和第 2 种形式 NAG 很相似, 本文主要研究第 2 种 NAG 和 AdaGrad 结合的问题. 主要贡献包括 3 个方面.

(1) 提出了一种 AdaNAG 算法. 所提算法保持了 AdaGrad 自适应策略累积历史平方梯度信息的特点, 克服了 AcceleGrad 对所有维度数据无法区别对待的缺陷;

(2) 针对约束的非光滑凸优化问题, 证明了本文提出的 AdaNAG 具有 $O(1/\sqrt{t})$ 的最优个体收敛速率(见定理 1). 据我们所知: 这一结果填补了带自适应策略 NAG 个体最优收敛性方面的缺失, 比 AcceleGrad 仅具有最优的加权平均收敛性结果要强;

(3) 本文选择了典型的 $l_1$ 范数约束下的 Hinge 损失函数分类和 $l_1$ 损失函数回归优化问题, 通过与几种具有最优收敛速率算法的比较, 实验验证了理论分析的正确性, 也表明了所提算法的性能优于 AcceleGrad.

## 1 相关工作

本文主要考虑求解如下非光滑约束优化问题:

$$\left.\begin{aligned}&\min f(\boldsymbol{w})\\&\text{s.t. } \boldsymbol{w}\in Q\end{aligned}\right\} \tag{1}$$

其中, $Q\in\mathbb{R}^d$ 是闭凸集合, 且 $f$ 是 $Q$ 上的非光滑凸函数. 令 $\boldsymbol{w}^*=\arg\min_{\boldsymbol{w}\in Q} f(\boldsymbol{w})$ 为公式(1)的一个最优解.

根据算法的输出方式, 我们将收敛速率分为平均收敛速率和个体收敛速率, 平均收敛速率的表达式为

$$f(\bar{\boldsymbol{w}}_t)-f(\boldsymbol{w}^*) \tag{2}$$

其中, $\bar{\boldsymbol{w}}_t=\frac{1}{t}\sum_{k=1}^{t}\boldsymbol{w}_k$. 当算法以单步个体的方式输出时有个体收敛速率, 其表达式为

$$f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*) \tag{3}$$

为解问题(1), 采用投影次梯度方法, 关键迭代步骤为

$$\boldsymbol{w}_{t+1}=\boldsymbol{P}_Q(\boldsymbol{w}_t-\alpha_t\nabla f(\boldsymbol{w}_t)) \tag{4}$$

其中, $\boldsymbol{P}_Q(\cdot)$表示集合 $Q$ 上的投影算子, $\nabla f(\boldsymbol{w}_t)$表示 $f$ 在 $\boldsymbol{w}_t$ 处的次梯度, $\alpha_t$ 为迭代步长.

当样本服从独立同分布时, 用次梯度的无偏估计$\nabla f_m(\boldsymbol{w}_t)$替换$\nabla f(\boldsymbol{w}_t)$, 则公式(4)转变为 SGD 算法. $\nabla f(\boldsymbol{w}_t)$决定了参数的更新方向, $\alpha_t$ 决定了参数的更新大小. 对于公式(1)所描述的非光滑凸优化问题, 当设置 $\alpha_t=\alpha/\sqrt{t}$, $\alpha>0$ 时, SGD 最优的平均收敛速率是 $O(1/\sqrt{t})$. 但其个体收敛速率较平均收敛速率更难获得, 文献[4−6]证明了 SGD 最优的个体收敛速率是 $O(\log t/\sqrt{t})$.

AdaGrad 的关键迭代步骤如下:

$$\begin{cases}\boldsymbol{v}_t=\boldsymbol{v}_{t-1}+\boldsymbol{g}_t\odot\boldsymbol{g}_t\\ V_t=diag(\sqrt{\boldsymbol{v}_t})+\varepsilon I\\ \boldsymbol{w}_{t+1}=\boldsymbol{P}_Q^{V_t}(\boldsymbol{w}_t-\alpha V_t^{-1}\boldsymbol{g}_t)\end{cases}\tag{5}$$

其中, $\boldsymbol{g}_t=\nabla f(\boldsymbol{w}_t)$; $\odot$表示元素积; $I$ 为单位对称矩阵; $\varepsilon$为平滑系数, 通常取值 1e–8; $\alpha$为某一固定超参数; $\boldsymbol{P}_Q^{V_t}(\boldsymbol{u})=\arg\min_{\boldsymbol{w}\in Q}\|\boldsymbol{w}-\boldsymbol{u}\|_{V_t}^2$ 代表加权投影算子, $V_t^{-1}$ 为矩阵 $V_t$ 的逆.

由公式(5)可知, $V_t^{-1}$ 为对角矩阵, AdaGrad 算法因此可视为用带矩阵的自适应步长 $\alpha V_t^{-1}$ 替换公式(4)中的步长$\alpha_t$而产生. $V_t^{-1}$ 累积历史平方梯度信息, 其对角线上数值对应各维度参数的更新权重, 从而使步长在不同维度上体现差异性.

注意, $v_{t,i}=\sum_{k=1}^{t}g_{k,i}^2$, 用 $A_{t,i}$ 表示矩阵 $A_t$ 的第 $i$ 维分量值, 则 AdaGrad 自适应步长第 $i$ 维元素可表示如下:

$$\alpha V_{t,i}^{-1}=\frac{\alpha}{\sqrt{t}}\frac{1}{\sqrt{\frac{1}{t}\sum_{k=1}^{t}g_{k,i}^2}+\frac{\varepsilon}{\sqrt{t}}}\tag{6}$$

由上式可知, AdaGrad 的自适应步长策略分为 3 个部分: 第 1 部分为与 SGD 相同阶的步长 $\alpha/\sqrt{t}$, 第 2 部分为分母上以算数平均的方式累积历史平方梯度信息 $\left(\frac{1}{t}\sum_{k=1}^{t}g_{k,i}^2\right)$, 第 3 部分为平滑项 $\varepsilon/\sqrt{t}$. 由于第 2 项和第 3 项组成部分的数量阶仍为常数阶的, AdaGrad 的迭代步长 $\alpha V_t^{-1}$ 最终还是满足 $O(1/\sqrt{t})$. 数据某一维度变化量很小时步长矩阵对应维度值很小, 迭代步长变大从而加快收敛, 相反在数据出现频率高的维度上降低参数更新速度. 特别对于稀疏数据, 大部分维度变化为 0, 因此收敛速度更快.

另一方面, 当公式(1)中 $f$ 为光滑凸函数时, 公式(4)的等价形式如下:

$$\boldsymbol{w}_{t+1}=\arg\min_{\boldsymbol{w}\in Q}\left\{\langle\nabla f(\boldsymbol{w}_t),\boldsymbol{w}\rangle+\frac{L}{2}\|\boldsymbol{w}-\boldsymbol{w}_t\|^2\right\}\tag{7}$$

其中, $L$ 为函数梯度 Lipschitz 常数, 转为公式(4)可看出步长$\alpha_t=1/L$. 针对公式(7), 文献[13]证明了有以下结论成立:

$$f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)\leqslant\frac{L}{2t}\|\boldsymbol{w}-\boldsymbol{w}_0\|^2\tag{8}$$

上式表明: 投影次梯度在光滑情况下, 取常数步长能够达到 $O(1/t)$的个体收敛速率.

NAG 动量方法早期集中研究于光滑凸优化问题中, 文献[13]中整理了 NAG 方法在无约束情形中欧几里得空间下可等价转换的 3 种形式, 其中, 第 1 种形式表示如下:

$$\begin{cases}\boldsymbol{y}_t=\boldsymbol{w}_t+\theta_t(\theta_{t-1}^{-1}-1)(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})\\ \boldsymbol{w}_{t+1}=\boldsymbol{P}_Q(\boldsymbol{y}_t-\nabla f(\boldsymbol{y}_t)/L)\\ \theta_{t+1}=(\sqrt{\theta_t^4+4\theta_t^2}-\theta_t^2)/2\end{cases}\tag{9}$$

其中, $\boldsymbol{w}_0=\boldsymbol{w}_{-1}\in Q$, $\theta_0=\theta_{-1}=1$. 可以看到: NAG 方法将承载动量特性的$\boldsymbol{y}_t$引入迭代步骤, 将参数更新方向由$\nabla f(\boldsymbol{w}_t)$变更为$\nabla f(\boldsymbol{y}_t)$. 使用这样的加速策略, 在处理光滑优化问题时, 可以将投影梯度算法收敛速率提升一个数量级至 $O(1/t^2)$. 第 2 种形式的 NAG 方法描述如下:

$$\begin{cases}\boldsymbol{y}_t=(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{z}_t\\ \boldsymbol{z}_{t+1}=\boldsymbol{P}_Q(\boldsymbol{z}_t-\nabla f(\boldsymbol{y}_t)/L\theta_t)\\ \boldsymbol{w}_{t+1}=(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{z}_{t+1}\\ \theta_{t+1}=(\sqrt{\theta_t^4+4\theta_t^2}-\theta_t^2)/2\end{cases}\tag{10}$$

其中, $\boldsymbol{w}_0, \boldsymbol{z}_0\in Q$, $\theta_0=\theta_{-1}=1$. 注意: 公式(9)中, $\boldsymbol{y}_t$可能会移动到约束区域外导致不便于收敛性分析; 但公式(10)中由于 $0<\theta_k<1$, 因此 $\boldsymbol{y}_t$始终在约束区域内. 同时, 从公式(10)第 2 行可以很方便看出, 所用步长为$\alpha_t=1/L\theta_t$.

AcceleGrad 算法具体描述如下:

$$\begin{cases} \boldsymbol{x}_{t+1}=\dfrac{1}{\theta_t}\boldsymbol{y}_t+\left(1-\dfrac{1}{\theta_t}\right)\boldsymbol{w}_t \\ \boldsymbol{y}_{t+1}=P_Q(\boldsymbol{y}_t-\theta_t\eta_t\nabla f(\boldsymbol{x}_{t+1})) \\ \boldsymbol{w}_{t+1}=\boldsymbol{x}_{t+1}-\eta_t\nabla f(\boldsymbol{x}_{t+1}) \\ \bar{\boldsymbol{w}}_{t+1}=\sum_{k=0}^{t}\theta_k\boldsymbol{w}_{k+1} \end{cases} \tag{11}$$

其中, 参数$\theta_t$取值为

$$\theta_t=\begin{cases} 1, & 0\leqslant t\leqslant 2 \\ \dfrac{1}{4}(t+1), & t\geqslant 3 \end{cases} \tag{12}$$

参数$\eta_t$设置为

$$\eta_t=\frac{2D}{(M^2+\sum_{k=0}^{t}\theta_k^2\,\|\nabla f(\boldsymbol{x}_{k+1})\|^2)^{1/2}} \tag{13}$$

其中, $D:=\max\limits_{\boldsymbol{w},\boldsymbol{u}\in Q}\|\boldsymbol{w}-\boldsymbol{u}\|^2$, $M$ 表示梯度$\nabla f(\boldsymbol{x}_{k+1})$范数的上界.

从公式(11)第 2 行可以看出, AcceleGrad 算法本质上迭代步长为

$$\theta_t\eta_t=\frac{2D\theta_t}{(M^2+\sum_{k=0}^{t}\theta_k^2\,\|\nabla f(\boldsymbol{x}_{k+1})\|^2)^{1/2}} \tag{14}$$

从上式可以看出: AcceleGrad 的自适应步长策略与 AdaGrad 风格不同, 需要额外添加梯度范数的上界 $M$、约束区域的最大直径 $D$ 这两个超参数. 引入一个条件参数$\theta_t$作为累积梯度历史信息的权重, 可以强调最新的梯度信息. 虽然 AcceleGrad 没有使用矩阵而是累积了历史梯度范数的平方项, 但是对于非光滑凸问题, 其步长仍然符合常规的 $O(1/\sqrt{t})$. 组合上述所有参数得到步长, 原因之一是为了搭建统一的算法框架, 使其无论在处理光滑或非光滑问题时, 都无需获取光滑先验信息或修改算法保证收敛性. 但却失去了 AdaGrad 自适应步长策略能体现不同维度差异的特性.

## 2 AdaNAG 方法

对于非光滑凸优化问题, 为了在第 2 种形式 NAG 中引入自适应策略, 我们的思路是: 摒弃针对光滑优化问题步长策略$\alpha_t$=1/$L\theta_t$的步长设置, 取而代之的是, 采用类似公式(5)的自适应策略. 本文提出的AdaNAG方法具体描述如下:

$$\begin{cases} \boldsymbol{y}_t=(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{z}_t \\ \boldsymbol{v}_t=\boldsymbol{v}_{t-1}+(\hat{\boldsymbol{g}}_t\odot\hat{\boldsymbol{g}}_t)+\delta \\ V_t=diag(\sqrt{\boldsymbol{v}_t}) \\ \boldsymbol{z}_{t+1}=\boldsymbol{P}_Q^{V_t}(\boldsymbol{z}_t-\eta_t\theta_t^{-1}V_t^{-1}\hat{\boldsymbol{g}}_t) \\ \boldsymbol{w}_{t+1}=(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{z}_{t+1} \\ \theta_{t+1}=(\sqrt{\theta_t^4+4\theta_t^2}-\theta_t^2)/2 \end{cases} \tag{15}$$

其中, $\boldsymbol{w}_0\in Q$; $\boldsymbol{z}_0\in Q$; $\theta_0=\theta_{-1}=1$; $\hat{\boldsymbol{g}}_t=\nabla f(\boldsymbol{y}_t)$; $V_t$是半正定的对角矩阵; $\delta$为平滑系数, 取值为 1e–12. 注意:

$$\theta_{t+1}=(\sqrt{\theta_t^4+4\theta_t^2}-\theta_t^2)/2 \tag{16}$$

等价于:

$$\frac{1}{\theta_{t-1}^2}=\frac{1-\theta_t}{\theta_t^2} \tag{17}$$

再通过简单归纳论证, 可证明, 对$\forall t$=0,1,2,…, 有下式存在:

$$\theta_t\leqslant 2/(t+2) \tag{18}$$

特别地, 我们取时变动量参数和时变步长参数如下:

$$\theta_t=2/(t+2),\ \eta_t=\eta/(t+2),\ \eta>0 \tag{19}$$

从公式(15)第 4 行可以看出, 本文提出的 AdaNAG 方法的迭代步长为 $\eta_t\theta_t^{-1}V_t^{-1}$. 注意, $v_{t,i}=\sum_{k=1}^{t}g_{k,i}^2+t\delta$, 考虑步长第 $i$ 维元素可表示如下:

$$\eta_t\theta_t^{-1}V_{t,i}^{-1}=\frac{\eta}{2\sqrt{t}}\frac{1}{\sqrt{\frac{1}{t}\sum_{k=1}^{t}g_{k,i}^2+\delta}}.$$

从上式可以清楚地观察到: AdaNAG 方法的有效步长为 $\eta/2\sqrt{t}$, 算法整体迭代步长符合一阶梯度算法 $O(1/\sqrt{t})$ 的步长设置. 与 AcceleGrad 算法累积历史梯度范数的平方项不同, AdaNAG 继承 AdaGrad 步长风格, 用 $\sqrt{\frac{1}{t}\sum_{k=1}^{t}g_{k,i}^2+\delta}$ 来积累矩阵各维度数值, 来对不同维度步长进行加权区分, 从而体现待训参数之间迭代步长的差异性.

## 3 AdaNAG 方法个体收敛速率分析

AdaNAG 的个体收敛速率证明思路与文献[13]中第 2 种形式 NAG 方法类似, 不同的是, 需要处理因添加自适应矩阵带来的额外项. 具体证明可见定理 1, 先由函数 $f$ 关于常数 $M$ 满足 Lipschitz 连续条件, 得到 $|f(\boldsymbol{w}_{t+1})-f(\boldsymbol{y}_t)|\leqslant M\|\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\|$; 再将之与引理 1、引理 2 以及凸函数性质结合, 得到形如 $f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w}^*)\leqslant f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)+X$ 的式子, 递归后得形如 $f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w}^*)\leqslant Y+Z$ 的式子; 最后由引理 3 和引理 4 分别处理因递归产生的累加项 $Y$ 和 $Z$. 下面我们需要给出一些假设条件, 这些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:

$$\|\hat{\boldsymbol{g}}_t\|_\infty\leqslant M.$$

**假设 2(约束区域有界)**. 存在一个常数 $D>0$, 使得:

$$\max_{\forall \boldsymbol{w},\boldsymbol{z}\in Q}\|\boldsymbol{w}-\boldsymbol{z}\|_\infty\leqslant D.$$

**引理 1(三点性质)**. 假设 $\psi:\xi\to(-\infty,+\infty]$为下半连续的凸函数, 其中, $\xi$是具有范数$\|\cdot\|$的有限维实线性空间. 对任意的 $\boldsymbol{w}, \boldsymbol{z}_t\in Q$, $D(\boldsymbol{w},\boldsymbol{z}_t)$为强凸函数, 并满足:

$$\boldsymbol{z}_{t+1}=\arg\min_{\boldsymbol{w}\in Q}\{\psi(\boldsymbol{w})+D(\boldsymbol{w},\boldsymbol{z}_t)\},$$

则有以下不等式成立:

$$\psi(\boldsymbol{w})+D(\boldsymbol{w},\boldsymbol{z}_t)\geqslant\psi(\boldsymbol{z}_{t+1})+D(\boldsymbol{z}_{t+1},\boldsymbol{z}_t)+D(\boldsymbol{w},\boldsymbol{z}_{t+1}).$$

具体证明见文献[26]中引理 3.2.

**引理 2**. 令 $\{\boldsymbol{w}_t\}_{t=1}^{\infty}$, $\{\boldsymbol{y}_t\}_{t=1}^{\infty}$, $\{\boldsymbol{z}_t\}_{t=1}^{\infty}$ 和 $\{V_t\}_{t=1}^{\infty}$ 由公式(15)产生, $M$ 由假设 1 得到, $\theta_t$ 和 $\eta_t$ 在公式(19)中定义, 则下式成立:

$$M\|\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\|=\sum_{i=1}^{d}\frac{\eta_t M^2}{2V_{t,i}}+\frac{\theta_t^2}{2\eta_t}\|\boldsymbol{z}_t-\boldsymbol{z}_{t+1}\|_{V_t}^2.$$

具体证明见附录 1.

**引理 3**. 令 $\{\boldsymbol{w}_t\}_{t=1}^{\infty}$, $\{\boldsymbol{z}_t\}_{t=1}^{\infty}$ 和 $\{V_t\}_{t=1}^{\infty}$ 由公式(15)产生, $M$ 由假设 1 得到, $D$ 由假设 2 得到, $\theta_t$ 和 $\eta_t$ 在公式(19)中定义, 则下式成立:

$$\sum_{k=1}^{t}\left[\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_k\|_{V_k}^2-\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_{k+1}\|_{V_k}^2\right]\leqslant\frac{dD^2\sqrt{(M^2+\delta)}(t+2)\sqrt{t}}{2\eta}.$$

具体证明见附录 2.

**引理 4**. 令 $\{\boldsymbol{w}_t\}_{t=1}^{\infty}$, $\{\boldsymbol{y}_t\}_{t=1}^{\infty}$ 和 $\{\boldsymbol{z}_t\}_{t=1}^{\infty}$ 由公式(15)产生, $M$ 由假设 1 得到, $\theta_t$ 和 $\eta_t$ 在公式(19)中定义, 则下式成立:

$$\sum_{k=1}^{t}\frac{1}{\theta_k^2}\sum_{i=1}^{d}\frac{\eta_k M^2}{2V_{k,i}}\leqslant\frac{\eta dM^2(t+2)\sqrt{t}}{8\sqrt{\delta}}.$$

具体证明见附录 3.

**定理 1**. 令 $\{\boldsymbol{w}_t\}_{t=1}^{\infty}$, $\{\boldsymbol{y}_t\}_{t=1}^{\infty}$, $\{\boldsymbol{z}_t\}_{t=1}^{\infty}$ 和 $\{V_t\}_{t=1}^{\infty}$ 由公式(15)产生, $M$ 由假设 1 得到, $D$ 由假设 2 得到, $\boldsymbol{w}^*\in Q$ 为问题(1)的一个最优解, $\theta_t$ 和 $\eta_t$ 在公式(19)中定义, 结合引理 1引理 4, 下式成立:

$$f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w}^*)\leqslant\frac{\eta dM^2\sqrt{t}}{2\sqrt{\delta}(t+2)}+\frac{2dD^2\sqrt{(M^2+\delta)}\sqrt{t}}{\eta(t+2)}.$$

上式表明: AdaNAG 具有 $O(1/\sqrt{t})$ 的个体收敛速率, 相比 SGD 能达到的 $O(\log t/\sqrt{t})$ 最优个体收敛速率, 体现出了 NAG 动量方法的加速性. 具体证明见附录 4.

## 4 实验与结果

本节分别在分类和回归问题中, 对 AdaNAG 算法个体收敛速率的理论分析进行实验验证. 两组实验均采用 6 种优化算法进行比较, 这些方法分别为个体形式输出的 Frank Wolfe 算法[27]、加权平均形式输出的 AcceleGrad 算法[19]、加权平均形式输出的 UniXGrad 算法[22]、个体形式输出的 NAG 方法[16]、平均形式输出的 AdaGrad 算法[7]以及个体形式输出的 AdaNAG 算法. 从理论分析的角度来说, 这几种优化方法的收敛速率均达到了最优. 但在处理高维度数据时, AdaNAG 应该比 AcceleGrad 和 UniXGrad 收敛更快.

两组实验均参照文献[28]的解约束优化问题, 其中, 投影约束域 $Q$ 为 $l_1$ 范数球$\{\|\boldsymbol{w}\|_1\leqslant z\}$, 采用 SLEP 工具箱[29]中的函数 eplb 实现该投影运算, 其原理是, 通过二分法求得精确解. 根据数据集的不同, 约束参数 $z$ 对应选取不同的值, 但同一数据集中各算法取相同的 $z$. Frank Wolfe 算法本身就是解约束问题的算法, 因此不需要投影算子.

根据各算法收敛性结论确定步长超参数的设置, NAG 方法取$\alpha$=0.1, AdaNAG 取$\eta$=0.1, AdaGrad 取$\alpha$=0.01, AcceleGrad 和 UniXGrad 分别遵循文献[19,22]中的设置, Frank Wolfe 算法无超参数. 所有算法在每个数据集上运行 10 次, 并取平均值绘制收敛曲线图.

### 4.1 标准数据集上分类实验结果与分析

实验采用 6 个标准数据集, 这些数据集均来自于 LIBSVM 网站, 具体描述见表 1.

表 1 标准数据集描述

| 数据集 | 训练样本数 | 维数 |
|---|---|---|
| phishing | 11 055 | 68 |
| mushrooms | 8 124 | 112 |
| w8a | 49 749 | 300 |
| protein | 17 766 | 357 |
| sector | 6 412 | 55 197 |
| news20 | 15 935 | 62 061 |

选择典型非光滑 hinge 损失作为目标函数, 具体描述如下:

$$\min\ \frac{1}{M}\sum_{(\boldsymbol{x},y)\in S}\max\{0,1-y\langle\boldsymbol{w},\boldsymbol{x}\rangle\}$$
$$\text{s.t.}\ \ \|\boldsymbol{w}\|_1\leqslant z$$

其中, $S=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{M}$ 为全体样本集, $\boldsymbol{x}_i\in\mathbb{R}^d$, $y_i\in\{-1,1\}$.

根据文献[30]中批处理向随机转换技巧, 算法(15)可以推广到随机形式, 因此次梯度的计算方式如下:

$$\nabla f_m(\boldsymbol{w}_t)=\frac{1}{k}\sum_{(\boldsymbol{x}_m,y_m)\in A_t^+}y_m\boldsymbol{x}_m,$$

其中, $\boldsymbol{x}_m$ 代表第 $m$ 个样本的特征, $y_m$ 代表第 $m$ 个样本的标签, $A_t\subseteq S$ 且$|A_t|=k$, $A_t^+=\{(\boldsymbol{x}_m,y_m)\in A_t:y_m\langle\boldsymbol{w}_t,\boldsymbol{x}_m\rangle<1\}$. 实验中, $k$ 取值为 1, 也就是每次更新参数时只用一个样本计算次梯度.

图 1 为 6 种算法的收敛速率对比图, 纵坐标表示当前目标函数值与目标函数最优值之差(最优值取所有迭代结果中的最小值), 横坐标表示迭代次数. 紫红色圆圈标记的实线代表个体形式输出的 Frank Wolfe 算法的收敛趋势, 绿色星号标记的点虚线代表加权平均形式输出的 AcceleGrad 算法的收敛趋势, 深蓝色叉号标记的实线代表加权平均形式输出的 UniXGrad 算法的收敛趋势, 红色方框标记的点划线代表本文提出的个体形式输出的 AdaNAG 算法的收敛趋势, 青绿色菱形标记的实线代表个体形式输出的 NAG 方法收敛趋势, 黑色三角形标记的点虚线代表平均形式输出的 AdaGrad 算法的收敛趋势. 从图 1 可以看出: 6 种算法在 6 个标准数据集上迭代 10 000 步之后, 都至少达到 $10^{-2}$ 的精度, 可以说均表现出基本相同的收敛趋势; 并且在同一精度要求下, AdaNAG 的收敛速度总体上是最快的. 这与理论分析的结果相吻合.

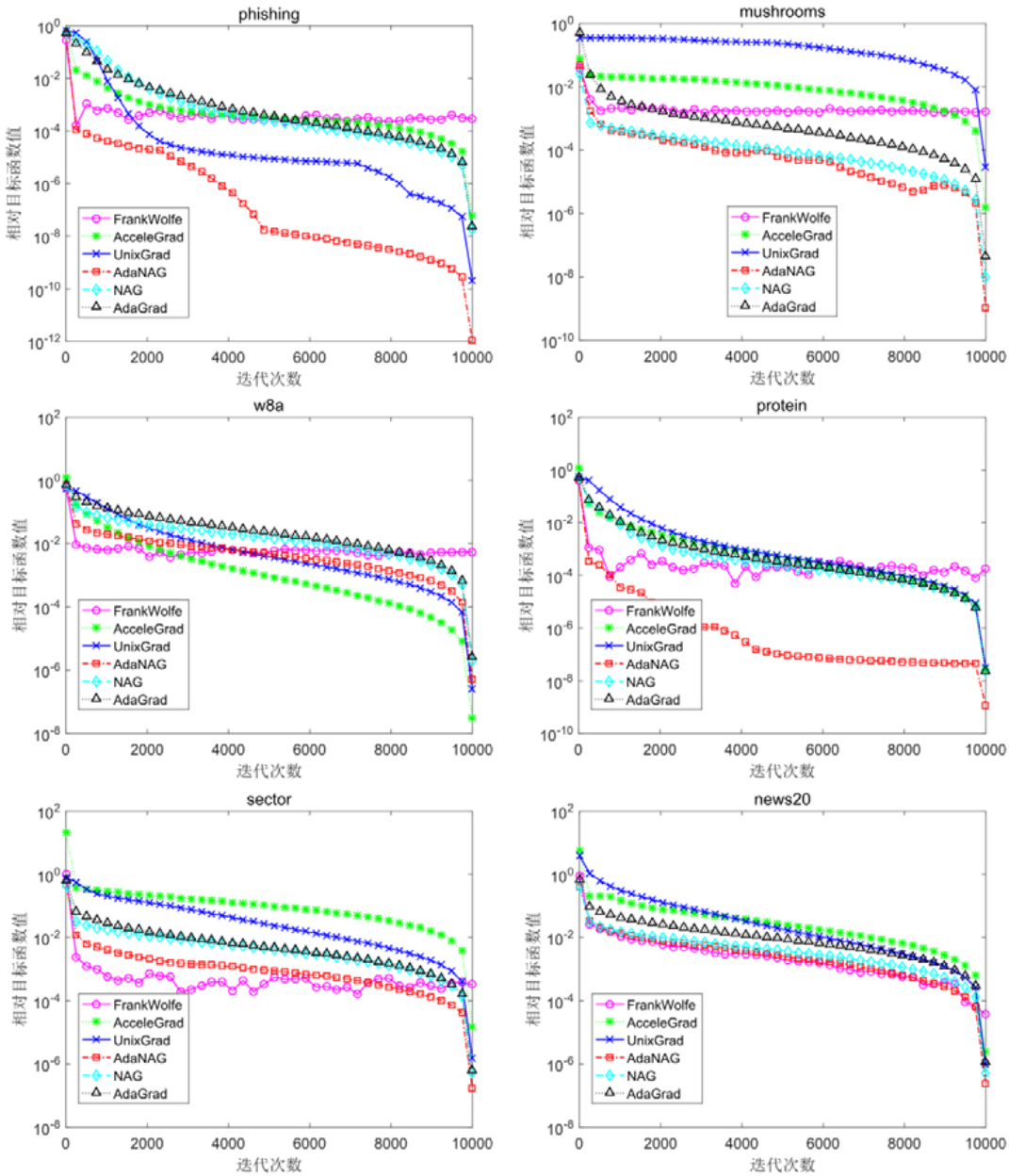



图 1　标准数据集上收敛速率比较图(迭代次数)

另外, 从图 1 中还可以看出: 对于维数较高的后两个数据库, AdaNAG 比 AcceleGrad 和 UniXGrad 收敛效果更好. 这是由于采用了不同类型的自适应步长策略所导致的. AcceleGrad 和 UniXGrad 在步长策略中都有对梯度求范数的操作, 没有体现出步长在不同维度上的差异性. 反观 AdaNAG, 基于 AdaGrad 类型自适应步长策略适合处理高维度数据集.

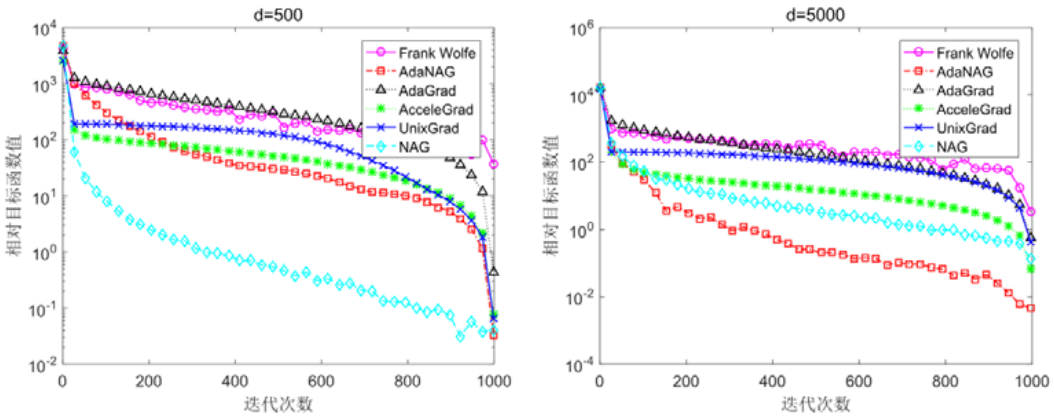
### 4.2 人工数据集上回归实验结果与分析

为进一步说明所提算法在非光滑凸优化问题中的应用, 我们添加了和文献[19]相同的回归实验进行了对比. 具体来说, 目标函数如下:

$$\min\ f(\boldsymbol{w}) = \|A\boldsymbol{w}-\boldsymbol{b}\|_1$$
$$\text{s.t.}\ \ \|\boldsymbol{w}\|_1 \leqslant z$$

其中, $A\in\mathbb{R}^{n\times d}$ 是随机生成的高斯矩阵; $\boldsymbol{b}=A\boldsymbol{w}^{\dagger}+\boldsymbol{\rho}$, $\boldsymbol{w}^{\dagger}$ 是 $d$ 维稀疏向量, 且前 50 维元素为±1; $\boldsymbol{\rho}$是高斯噪声.

6 种算法的收敛速率对比如图 2 所示, 其中, 横坐标代表迭代次数, 纵坐标代表当前目标函数值与目标函数最优值之差(最优值取所有迭代结果中的最小值), 图中曲线和算法的对应关系与上文一致. 当设置参数 $n$=200, $d$=500 时, 结果见图 2(a), 可以看出, AdaNAG 的表现优于 AcceleGrad 和 UniXGrad. 设置 $n$=200, $d$=5000 的实验结果见图 2(b), 此时, AdaNAG 性能超过其他所有比较算法; 同时也表明, 所提 AdaNAG 算法适合处理高维稀疏数据.




(a) (b)

图 2 人工数据集上收敛速率比较图(迭代次数)

## 5 结 论

本文提出了一种名为 AdaNAG 的自适应 Nesterov 加速梯度方法, 证明了 AdaNAG 算法能达到更好的 $O(1/\sqrt{t})$ 的个体收敛速率, 体现了 NAG 方法的加速性. 据我们所知, 这是第一个被证明具有最优个体收敛速率的自适应步长策略与 NAG 方法结合的算法. 与 AcceleGrad 算法相比, AdaNAG 继承了 AdaGrad 的自适应风格, 步长使用了矩阵进行加权, 体现出了不同维度的差异性. 实验验证了所提算法在解决非光滑凸优化问题时比 AcceleGrad 更优. 下一步, 我们将继续研究自适应步长策略下, NAG 方法在强凸情形下的个体收敛性及其在深度学习中的应用.

**References**:

[1] Zinkevich M. Online convex programming and generalized infinitesimal gradient ascent. In: Proc. of the 20th Int'l Conf. on Machine Learning (ICML 2003). New York: Association for Computing Machinery, 2003. 928−936.

[2] Shalev-Shwartz S, Singer Y, Srebro N. Pegasos: Primal estimated sub-GrAdientSOlver for SVM. In: Proc. of the 24th Int'l Conf. on Machine Learning (ICML 2007). New York: Association for Computing Machinery, 2007. 807−814.

[3] Shamir O, Zhang T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In: Proc. of the 30th Int'l Conf. on Machine Learning (ICML 2013). New York: Association for Computing Machinery, 2013. 71−79.

[4] Ge R, Jain P, Kakade SM, *et al*. Open problem: Do good algorithms necessarily query bad points? In: Proc. of the 32nd Annual Conf. on Learning Theory (COLT 2019). 2019. 3190−3193.

[5] Harvey NJA, Plan Y, Randhawa S. Tight analyses for non-smooth stochastic gradient descent. In: Proc. of the 32nd Annual Conf. on Learning Theory (COLT 2019). 2019. 1579−1613.

[6] Jain P, Nagaraj D, Netrapalli P. Making the last iterate of SGD information theoretically optimal. In: Proc. of the 32nd Annual Conf. on Learning Theory (COLT 2019). 2019. 1752−1755.

[7] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 2011, 12(7): 257−269.

[8] Zeiler MD. ADADELTA: An adaptive learning rate method. arXiv: 1212.5701, 2012.

[9] Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations (ICLR 2015). 2015. 1−13.

[10] Polyak BT. Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics, 1964, 4(5): 1−17.

[11] Nesterov Y. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. Soviet Mathematics Doklady, 1983, 27(2): 372−376.

[12] Nemirovsky AS, Yudin DB. Problem Complexity and Method Efficiency in optimization. New York: Wiley-Interscience, 1983.

[13] Tseng P. Approximation accuracy, gradient methods, and error bound for structured convex optimization. Mathematical Programming, 2010, 125(2): 263−295.

[14] Lan G. An optimal method for stochastic composite optimization. Mathematical Programming, 2012, 133(1-2): 365−397.

[15] Devolder O, Glineur F, Nesterov Y. First-Order methods of smooth convex optimization with inexact oracle. Mathematical Programming, 2014, 146(1-2): 37−75.

[16] Tao W, Pan Z, Wu G, *et al*. The stength of Nestrerov's extrapolation in the individual convergence of nonsmooth optimization. IEEE Trans. on Neural Networks and Learning Systems, 2020, 31(7): 2557−2568.

[17] Liu YX, Cheng YJ, Tao Q. Individual convergence of NAG with biased gradient in nonsmooth cases. Ruan Jian Xue Bao/Journal of Software, 2020, 31(4): 1051−1062 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5926.htm [doi: 10.13328/j.cnki.jos.005926]

[18] Dozat T. Incorporating Nesterov momentum into Adam. In: Proc. of the 4th Int'l Conf. on Learning Representations (ICLR 2016). 2016.

[19] Levy KY, Yurtsever A, Cevher V. Online adaptive methods, universality and acceleration. In: Proc. of the 32nd Annual Conf. on Neural Information Processing Systems (NeurIPS 2018). Cambridge: MIT, 2018. 6501−6510.

[20] Levy KY. Online to offline conversions, universality and adaptive minibatch sizes. In: Proc. of the 31st Annual Conf. on Neural Information Processing Systems (NeurIPS 2017). Cambridge: MIT, 2017. 1613−1622.

[21] Allen-Zhu Z, Orecchia L. Linear coupling: An ultimate unification of gradient and mirror descent. In: Proc. of the 8th Innovations in Theoretical Computer Science Conf. (ITCS 2017). 2017. 3:1−3:22.

[22] Kavis A, Levy KY, Bach F, *et al*. UniXGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. In: Proc. of the 33th Annual Conf. on Neural Information Processing Systems (NeurIPS 2019). Cambridge: MIT, 2019. 6257−6266.

[23] Deng Q, Cheng Y, Lan G. Optimal adaptive and accelerated stochastic gradient descent. arXiv: 1810.00553, 2018.

[24] Tao W, Long S, Wu G, *et al*. The role of momentum parameters in the optimal convergence of adaptive Polyak's Heavy-Ball methods. In: Proc. of the 9th Int'l Conf. on Learning Representations (ICLR 2021). 2021.

[25] Cheng YJ, Tao W, Liu YX, *et al*. Optimal individual convergence rate of the Heavy-ball-based momentum methods. Journal of Computer Research and Development, 2019, 56(8): 1686−1694 (in Chinese with English abstract).

[26] Chen G, Teboulle M. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. SIAM Journal on Optimization, 1993, 3(3): 538−543. [doi: 10.1137/0803026]

[27] Jaggi M. Revisitting Frank-Wolfe: Projection-free sparse convex optimization. In: Proc. of the 30th Int'l Conf. on Machine Learning (ICML 2013). New York: Association for Computing Machinery, 2013. 427−435.

[28] Liu J, Ye J. Efficient euclidean projections in linear time. In: Proc. of the 26th Int'l Conf. on Machine Learning (ICML 2009). New York: Association for Computing Machinery, 2009. 657−664.

[29] Liu J, Ji S, Ye J. SLEP: Sparse learning with efficient projections. Arizona: Arizona State University, 2009.

[30] Rakhlin A, Shamir O, Sridharan K. Making gradient descent optimal for strongly convex stochastic optimization. In: Proc. of the 29th Int'l Conf. on Machine Learning (ICML 2012). New York: Association for Computing Machinery, 2012. 449−456.

[31] Chen X, Lin Q, Pena J. Optimal regularized dual averaging methods for stochastic optimization. In: Proc. of the 26th Annual Conf. on Neural Information Processing Systems (NIPS 2012). Cambridge: MIT, 2012. 404−412.

**附中文参考文献**:

[17] 刘宇翔, 程禹嘉, 陶卿. 梯度有偏情形非光滑问题 NAG 的个体收敛性. 软件学报, 2020, 31(4): 1051−1062. http://www.jos.org.cn/1000-9825/5926.htm [doi: 10.13328/j.cnki.jos.005926]

[25] 程禹嘉, 陶蔚, 刘宇翔, 陶卿. Heavy-Ball 型动量方法的最优个体收敛速率. 计算机研究与发展, 2019, 56(8): 1686−1694.

## 附录 1. 引理 2 证明

$$M\|\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\|=M\|\theta_t\boldsymbol{z}_t-\theta_t\boldsymbol{z}_{t+1}\|\leqslant\sum_{i=1}^{d}M\theta_t|z_{t,i}-z_{t+1,i}|\leqslant\sum_{i=1}^{d}\left[\frac{\eta_tM^2}{2V_{t,i}}+\frac{\theta_t^2V_{t,i}}{2\eta_t}|z_{t,i}-z_{t+1,i}|^2\right]=\sum_{i=1}^{d}\frac{\eta_tM^2}{2V_{t,i}}+\frac{\theta_t^2}{2\eta_t}\|\boldsymbol{z}_t-\boldsymbol{z}_{t+1}\|_{V_t}^2,$$

其中, 第 2 个不等式运用了杨氏不等式.

## 附录 2. 引理 3 证明

$$\begin{aligned}\sum_{k=1}^{t}\left[\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_k\|_{V_k}^2-\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_{k+1}\|_{V_k}^2\right]&=\frac{1}{2\eta_1}\|\boldsymbol{w}-\boldsymbol{z}_1\|_{V_1}^2-\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|_{V_t}^2+\sum_{k=2}^{t}\left[\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_k\|_{V_k}^2-\frac{1}{2\eta_{k-1}}\|\boldsymbol{w}-\boldsymbol{z}_k\|_{V_{k-1}}^2\right]\\&\leqslant\sum_{i=1}^{d}\frac{V_{1,i}}{2\eta_1}(w_i-z_{1,i})^2+\sum_{i=1}^{d}\sum_{k=2}^{t}\left(\frac{V_{k,i}}{2\eta_k}-\frac{V_{k-1,i}}{2\eta_{k-1}}\right)(w_i-z_{k,i})^2\\&\leqslant\sum_{i=1}^{d}\frac{V_{t,i}}{2\eta_t}D^2\\&=\sum_{i=1}^{d}\frac{\sqrt{\sum_{k=1}^{t}\hat{g}_{k,i}^2+t\delta}}{2\eta_t}D^2\\&\leqslant\sum_{i=1}^{d}\frac{\sqrt{t(M^2+\delta)}}{2\eta_t}D^2\\&=\frac{dD^2\sqrt{(M^2+\delta)}(t+2)\sqrt{t}}{2\eta},\end{aligned}$$

其中, 第 2 个不等式运用了假设 2.

**附录 3**. 引理 4 证明

$$\sum_{k=1}^{t}\frac{1}{\theta_k^2}\sum_{i=1}^{d}\frac{\eta_k M^2}{2V_{k,i}}=\frac{M^2}{2}\sum_{k=1}^{t}\frac{\eta_k}{\theta_k^2}\sum_{i=1}^{d}\frac{1}{V_{k,i}}=\frac{M^2}{2}\sum_{k=1}^{t}\frac{\eta_k}{\theta_k^2}\sum_{i=1}^{d}\frac{1}{\sqrt{\sum_{j=1}^{k}\hat{g}_{j,i}^2+k\delta}}\leqslant\frac{M^2}{2}\sum_{k=1}^{t}\frac{\eta_k}{\theta_k^2}\sum_{i=1}^{d}\frac{1}{\sqrt{k\delta}}\leqslant\frac{\eta d M^2(t+2)\sqrt{t}}{8\sqrt{\delta}}.$$

**附录 4**. 定理 1 证明

由假设 1 得知, $f$ 关于常数 $M$ 满足 Lipschitz 连续条件; 再根据文献[31]中的公式(2), 可以建立起 $f(\boldsymbol{w}_{t+1})$和 $f(\boldsymbol{y}_t)$之间的关系:

$$f(\boldsymbol{w}_{t+1})\leqslant f(\boldsymbol{y}_t)+\langle\hat{\boldsymbol{g}}_t,\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\rangle+M\|\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\|,$$

代入 $\boldsymbol{w}_{t+1}=(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{z}_{t+1}$, 有:

$$f(\boldsymbol{w}_{t+1})\leqslant f(\boldsymbol{y}_t)+\langle\hat{\boldsymbol{g}}_t,(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{z}_{t+1}-\boldsymbol{y}_t\rangle+M\|\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\|.$$

上式等价于:

$$f(\boldsymbol{w}_{t+1})\leqslant(1-\theta_t)[f(\boldsymbol{y}_t)+\langle\hat{\boldsymbol{g}}_t,\boldsymbol{w}_t-\boldsymbol{y}_t\rangle]+\theta_t[f(\boldsymbol{y}_t)+\langle\hat{\boldsymbol{g}}_t,z_{t+1}-\boldsymbol{y}_t\rangle]+M\|\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\|.$$

根据引理 2, 上式可得:

$$f(\boldsymbol{w}_{t+1})\leqslant(1-\theta_t)[f(\boldsymbol{y}_t)+\langle\hat{\boldsymbol{g}}_t,\boldsymbol{w}_t-\boldsymbol{y}_t\rangle]+\theta_t[f(\boldsymbol{y}_t)+\langle\hat{\boldsymbol{g}}_t,\boldsymbol{z}_{t+1}-\boldsymbol{y}_t\rangle]+\sum_{i=1}^{d}\frac{\eta_t M^2}{2V_{t,i}}+\frac{\theta_t^2}{2\eta_t}\|\boldsymbol{z}_t-\boldsymbol{z}_{t+1}\|_{V_t}^2 \tag{20}$$

注意, 公式(15)式的第 4 行等价于:

$$\boldsymbol{z}_{t+1}=\arg\min_{\boldsymbol{w}\in Q}\left\{f(\boldsymbol{y}_t)+\langle\hat{\boldsymbol{g}}_t,\boldsymbol{w}-\boldsymbol{y}_t\rangle+\frac{\theta_t}{2\eta_t}\|(\boldsymbol{w}-\boldsymbol{z}_t)\|_{V_t}^2\right\}.$$

根据三点性质得到:

$$f(\boldsymbol{y}_t)+\langle\hat{\boldsymbol{g}}_t,\boldsymbol{z}_{t+1}-\boldsymbol{y}_t\rangle+\frac{\theta_t}{2\eta_t}\|\boldsymbol{z}_{t+1}-\boldsymbol{z}_t\|_{V_t}^2+\frac{\theta_t}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|_{V_t}^2\leqslant f(\boldsymbol{y}_t)+\langle\hat{\boldsymbol{g}}_t,\boldsymbol{w}-\boldsymbol{y}_t\rangle+\frac{\theta_t}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_t\|_{V_t}^2.$$

代入公式(20), 得到:

$$f(\boldsymbol{w}_{t+1})\leqslant(1-\theta_t)[f(\boldsymbol{y}_t)+\langle\hat{\boldsymbol{g}}_t,\boldsymbol{w}_t-\boldsymbol{y}_t\rangle]+\theta_t[f(\boldsymbol{y}_t)+\langle\hat{\boldsymbol{g}}_t,\boldsymbol{w}-\boldsymbol{y}_t\rangle]+\frac{\theta_t^2}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_t\|_{V_t}^2-\frac{\theta_t^2}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|_{V_t}^2+\sum_{i=1}^{d}\frac{\eta_t M^2}{2V_{t,i}}.$$

联立凸函数性质:

$$f(\boldsymbol{y}_t)+\langle\hat{\boldsymbol{g}}_t,\boldsymbol{w}_t-\boldsymbol{y}_t\rangle\leqslant f(\boldsymbol{w}_t),\ f(\boldsymbol{y}_t)+\langle\hat{\boldsymbol{g}}_t,\boldsymbol{w}-\boldsymbol{y}_t\rangle\leqslant f(\boldsymbol{w}),$$

上式得:

$$f(\boldsymbol{w}_{t+1})\leqslant(1-\theta_t)f(\boldsymbol{w}_t)+\theta_t f(\boldsymbol{w})+\frac{\theta_t^2}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_t\|_{V_t}^2-\frac{\theta_t^2}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|_{V_t}^2+\sum_{i=1}^{d}\frac{\eta_t M^2}{2V_{t,i}}.$$

令 $e_t=f(\boldsymbol{w}_t)-f(\boldsymbol{w})$, 代入上式得:

$$e_{t+1}\leqslant(1-\theta_t)e_t+\sum_{i=1}^{d}\frac{\eta_t M^2}{2V_{t,i}}+\frac{\theta_t^2}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_t\|_{V_t}^2-\frac{\theta_t^2}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|_{V_t}^2.$$

两边同时除以 $\theta_t^2$, 得:

$$\frac{e_{t+1}}{\theta_t^2}\leqslant\frac{(1-\theta_t)e_t}{\theta_t^2}+\frac{1}{\theta_t^2}\sum_{i=1}^{d}\frac{\eta_t M^2}{2V_{t,i}}+\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_t\|_{V_t}^2-\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|_{V_t}^2.$$

注意, $\frac{1}{\theta_t^2}=\frac{1-\theta_{t+1}}{\theta_{t+1}^2}$. 上式等价于:

$$\frac{(1-\theta_{t+1})e_{t+1}}{\theta_{t+1}^2}\leqslant\frac{(1-\theta_t)e_t}{\theta_t^2}+\frac{1}{\theta_t^2}\sum_{i=1}^{d}\frac{\eta_t M^2}{2V_{t,i}}+\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_t\|_{V_t}^2-\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|_{V_t}^2.$$

递归运用上式, 得:

$$\frac{(1-\theta_{t+1})e_{t+1}}{\theta_{t+1}^2}\leqslant\frac{(1-\theta_0)e_0}{\theta_0^2}+\sum_{k=1}^{t}\frac{1}{\theta_k^2}\sum_{i=1}^{d}\frac{\eta_k M^2}{2V_{k,i}}+\sum_{k=1}^{t}\left[\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_k\|_{V_k}^2-\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_{k+1}\|_{V_k}^2\right].$$

将 $\theta_0$=1, $\frac{1}{\theta_t^2}=\frac{1-\theta_{t+1}}{\theta_{t+1}^2}$ 代入上式, 得:

$$\frac{e_{t+1}}{\theta_t^2}\leqslant\sum_{k=1}^{t}\frac{1}{\theta_k^2}\sum_{i=1}^{d}\frac{\eta_k M^2}{2V_{k,i}}+\sum_{k=1}^{t}\left[\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_k\|_{V_k}^2-\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_{k+1}\|_{V_k}^2\right].$$

上式联立引理 3 和引理 4, 可得:

$$\frac{e_{t+1}}{\theta_t^2}\leqslant\frac{\eta dM^2(t+2)\sqrt{t}}{8\sqrt{\delta}}+\frac{dD^2\sqrt{(M^2+\delta)}(t+2)\sqrt{t}}{2\eta}.$$

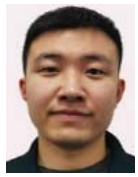
不等式两边同时乘 $\theta_t^2$, 得:

$$e_{t+1}=f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w})\leqslant\theta_t^2\left(\frac{\eta dM^2(t+2)\sqrt{t}}{8\sqrt{\delta}}+\frac{dD^2\sqrt{(M^2+\delta)}(t+2)\sqrt{t}}{2\eta}\right)=\frac{\eta dM^2\sqrt{t}}{2\sqrt{\delta}(t+2)}+\frac{2dD^2\sqrt{(M^2+\delta)}\sqrt{t}}{\eta(t+2)}.$$

令 $\boldsymbol{w}=\boldsymbol{w}^*$, 则定理 1 得证. □


**陇盛**(1998－), 男, 硕士生, 主要研究领域为机器学习, 模式识别.

**陶蔚**(1991－), 男, 博士, 助理研究员, 主要研究领域为机器学习.



**张泽东**(1994－), 男, 硕士生, 主要研究领域为机器学习, 模式识别.

**陶卿**(1965－), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为机器学习, 模式识别, 应用数学.