计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20190167
56(8):1686-1694,2019

Heavy-Ball 型动量方法的最优个体收敛速率

程禹嘉[1] 陶 蔚[2] 刘宇翔[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 合肥 230031)
[2](中国人民解放军陆军工程大学指挥控制工程学院 南京 210007)
(m13776555321@163.com)

Optimal Individual Convergence Rate of the Heavy-Ball-Based Momentum Methods

Cheng Yujia[1], Tao Wei[2], Liu Yuxiang[1], and Tao Qing[1]
[1](*Department of Information Engineering, Army Academy of Artillery and Air Defense of PLA, Hefei* 230031)
[2](*College of Command and Control Engineering, Army Engineering University of PLA, Nanjing* 210007)

Abstract The momentum method is widely used as an acceleration technique to improve the convergence of the first-order gradient algorithms. So far, the momentum methods discussed in most literatures are only limited to the accelerated method proposed by Nesterov, but the Heavy-ball momentum method proposed by Polyak is seldom studied. In particular, in the case of non-smooth objective functions, the individual optimal convergence of Nesterov accelerated methods has been derived, and it has high performance in solving sparse optimization problems. In contrast, while it has been proved that the Heavy-ball momentum method has an optimal convergence rate, it is only in terms of the averaged outputs. To our best knowledge, whether it has optimal individual convergence or not still remains unknown. In this paper, we focus on the non-smooth optimizations. We prove that the Heavy-ball momentum method achieves the optimal individual convergence by skillfully selecting the time-varying step-size, which indicates that Heavy-ball momentum is an efficient acceleration strategy for the individual convergence of the projected subgradient methods. As an application, the constrained hinge loss function optimization problems within an l_1-norm ball are considered. In comparison with other optimization algorithms, the experiments demonstrate the correctness of our theoretical analysis and performance of the proposed algorithms in keeping the sparsity.

Key words first-order gradient methods; momentum methods; individual convergence rate; heavy-ball methods; sparsity

摘 要 动量方法作为一种加速技巧被广泛用于提高一阶梯度优化算法的收敛速率.目前,大多数文献所讨论的动量方法仅限于 Nesterov 提出的加速方法,而对 Polyak 提出的 Heavy-ball 型动量方法的研究却较少.特别,在目标函数非光滑的情形下,Nesterov 加速方法具有最优的个体收敛性,并在稀疏优化问题的求解中具有很好的效果.但对于 Heavy-ball 型动量方法,目前仅仅获得了平均输出形式的最优收敛速率,个体收敛是否具有最优性仍然未知.对于非光滑优化问题,通过巧妙地设置步长,证明了 Heavy-ball 型动量方法具有最优的个体收敛速率,从而说明了 Heavy-ball 型动量方法可以将投影次梯度

收稿日期:2019-03-19;**修回日期**:2019-05-22
基金项目:国家自然科学基金项目(61673394)
This work was supported by the National Natural Science Foundation of China (61673394)
通信作者:陶卿(qing.tao@ia.ac.cn)

方法的个体收敛速率加速至最优.作为应用,考虑了 l_1 范数约束的 hinge 损失函数优化问题.通过与同类的优化算法相比,实验验证了该理论分析的正确性以及所提算法在保持稀疏性方面的良好性能.

关键词 一阶梯度方法;动量方法;个体收敛速率;Heavy-ball 方法;稀疏性

中图法分类号 TP181

机器学习问题普遍可以转化为求解目标函数最小值的优化问题,一阶梯度优化方法由于具有算法简单、迭代成本小等特点,成为处理大规模数据问题的首要选择.在此基础上发展起来的随机优化方法由于避免了每一次迭代都需要遍历整个样本集,充分利用训练样本集合的冗余性,从而具有计算代价低和实际收敛速率快等优点,已经成为处理大规模机器学习问题的实用方法[1].

动量方法是在经典梯度下降方法的基础上通过添加动量而获得的一种特殊的一阶优化方法.研究者将动量算法分为 2 类:一类是 Polyak 于 1964 年提出的 Heavy-ball 型动量方法[2],另一类是 1983 年 Nesterov 提出的 NAG(Nesterov accelerated gradient)型动量方法[3].这 2 类算法的主要差别在于所使用动量项的不同,前者只是使用了前一步的迭代信息而后者引入了当前步迭代算法的二阶信息.对于不同条件下的优化问题,这 2 类算法的收敛性也有不同的表现.当目标函数光滑时,NAG 具有最优的 $O(1/t^2)$收敛速率[3],其中 t 为算法的迭代步数.当目标函数强凸且二次可微时,尽管 Heavy-ball 型动量方法、梯度下降法和 NAG 方法都具有相同的线性收敛速率,但 Heavy-ball 型动量方法具有最小的收敛因子[2].随机动量方法被广泛应用于神经网络的训练,并显著的提高了其收敛性能[4].

NAG 是优化领域具有里程碑意义的算法,它填补了 Nemirovski 与 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[5]"的间隙,也吸引了众多机器学习研究者的关注.特别是针对大规模具有特定含义的正则化损失函数优化问题,研究工作层出不穷.早期重要的工作主要包括只要损失函数满足光滑性条件就可得到整个目标函数光滑时的最优收敛速率[6-7],以及 NAG 随机形式的最优收敛速率等等[8].近期 NAG 的研究主要集中在与其他优化方法的结合上.如 2015 年 Lin 等人基于 NAG 提出了一种通用的加速策略 Catalyst[9],在目标函数强凸的条件下将批处理优化方法、坐标优化方法和增量优化算法进行了加速.最近,Allen-Zhu 引入带有动量参数的方差项,提出了著名的 Katyusha 算法[10],成功地将方差减少方法与 NAG 相结合.2018 年 Shang 等人将 NAG 算法与 Mixed Optimization 算法相结合,仅使用了一个动量项就取得了与 Katyusha 算法相同的收敛速率[11-12].

与标准的梯度下降法相较,Heavy-ball 型动量方法在目标函数在某些方向变化较弱而在另一些方向变化很强的时候,可以取得好的加速效果,复杂度却几乎没有增加.但与 NAG 方法相比,Heavy-ball 型动量方法的研究却屈指可数.2014 年 Ghadimi 等人对 Heavy-ball 方法的收敛性进行了深入的研究,给出了目标函数光滑条件下的平均和个体收敛速率[13],但均未达到最优.2016 年 Yang 等人建立了一种含有多种参数的算法框架,统一处理了梯度下降法、Heavy-ball 方法以及 NAG 方法[14],在该框架中设置不同的参数即可得到不同的优化算法.这种统一的算法对于非光滑优化问题在平均输出方式下具有最优的收敛速率.

对于非光滑问题,目前大多数优化算法所获得的最优速率都基于加权平均和的输出形式,这种形式较易获得稳定的收敛性,但在结构优化特别是稀疏学习问题中[15-16],使用迭代过程中的个体输出能够获得比平均输出更好的稀疏效果.但个体解能否获得最优的收敛速率却显得困难重重,强凸条件下的个体收敛问题也成为了 open 问题[17].2013 年 Shamir 和 Zhang 提出了一种将 SGD 问题由平均输出方式得到个体收敛速率的技巧[18],成功得到了 $O(\log t/\sqrt{t})$的收敛速率,这是第一个关于 SGD 个体收敛速率的结果,但却与最优值相差一个对数因子.2015 年文献[19]采用线性差值技巧虽然保证了个体解最优的收敛性和稳定性,却由于插值的累积而失去了稀疏性.2018 年文献[20-21]将 NAG 步长策略引入到投影次梯度中,得到了最优个体收敛性并同时保证了良好的稀疏性.由于 Heavy-ball 方法与 NAG 在动量方法中具有同等重要的地位,Heavy-ball 方法是否也具有最优个体收敛性这一问题显然值得研究.

本文的主要工作有3个方面：

1) 对于非光滑优化问题，证明了Heavy-ball型动量方法具有最优的个体收敛速率.据我们所知，这一结果填补了Heavy-ball型动量方法在非光滑情形下个体最优收敛性研究的缺失，有助于更全面地理解Heavy-ball型动量方法，也说明了在处理非光滑问题时Heavy-ball型动量方法和NAG具有同样的重要地位.

2) 本文证明基于光滑情形下Heavy-ball型动量方法的收敛性分析[13]，但不同的是，为得到非光滑情形下的个体最优收敛速率，我们巧妙构造了步长和动量权重的迭代方式，同时利用Zinkevich在处理在线优化方法收敛性使用的技巧[22]，避免了变步长和权重导致的递归问题.

3) 通过典型的稀疏优化问题实验，验证了理论分析的正确性以及所提算法在保持稀疏性方面的良好性能.

1 典型动量方法的收敛性介绍

本节我们对2种动量方法的收敛性以及它们之间的联系和区别进行必要的介绍.

考虑有约束优化问题：

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

其中，$f(\boldsymbol{w})$为凸函数，$Q\subseteq\mathbb{R}^n$是有界闭凸集合，记$\boldsymbol{w}^*$为式(1)的一个最优解.批处理形式投影次梯度方法的迭代步骤为

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

$\boldsymbol{P}_Q$是Q上的投影算子[23-24]，$\boldsymbol{w}_t$为$\boldsymbol{w}$在第t步的输出，α_t为迭代步长，$\nabla f(\boldsymbol{w}_t)$是$f(\boldsymbol{w})$在$\boldsymbol{w}_t$处的次梯度.

对于形如式(2)的算法，所谓的平均收敛速率指的是$f(\hat{\boldsymbol{w}}_t)-f(\boldsymbol{w}^*)$的收敛速率，其中$\hat{\boldsymbol{w}}_t=\frac{1}{t}\sum_{k=1}^{t}\boldsymbol{w}_k$.而个体收敛速率指的是$f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)$的收敛速率.一般来说，特别是对非光滑优化问题，个体收敛更难获得最优速率[18].

Agarwal等人给出非光滑条件下式(2)的平均收敛速率为[25]

$$E[f(\hat{\boldsymbol{w}}_t)-f(\boldsymbol{w}^*)]\leqslant O(1/\sqrt{t}). \tag{3}$$

式(2)的个体收敛速率由Shamir和Zhang Tong证得[18]：

$$E[f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)]\leqslant O(\log(t)/\sqrt{t}), \tag{4}$$

这与最优速率之间还是存在着数量级上的差距.

Yang等人建立了随机梯度下降法与随机动量方法的统一框架[14]：

$$\begin{aligned}\boldsymbol{y}_{t+1}&=\boldsymbol{w}_t-\alpha\nabla f(\boldsymbol{w}_t)\\ \boldsymbol{y}_{t+1}^s&=\boldsymbol{w}_t-s\alpha\nabla f(\boldsymbol{w}_t)\\ \boldsymbol{w}_{t+1}&=\boldsymbol{y}_{t+1}+\beta(\boldsymbol{y}_{t+1}^s-\boldsymbol{y}_t^s),\end{aligned} \tag{5}$$

其中，β为动量参数，$s\geqslant 0$.随着s由大至小，式(5)依次变为

1) 当$s=\frac{1}{1-\beta}$时，为梯度下降法：

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\frac{\alpha}{1-\beta}\nabla f(\boldsymbol{w}_t); \tag{6}$$

2) 当$s=1$时，即为NAG方法：

$$\begin{aligned}\boldsymbol{y}_{t+1}&=\boldsymbol{w}_t-\alpha\nabla f(\boldsymbol{w}_t),\\ \boldsymbol{w}_{t+1}&=\boldsymbol{y}_{t+1}+\beta(\boldsymbol{y}_{t+1}-\boldsymbol{y}_t);\end{aligned} \tag{7}$$

3) 当$s=0$时，即为Heavy-ball方法：

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\alpha\nabla f(\boldsymbol{w}_t)+\beta(\boldsymbol{w}_t-\boldsymbol{w}_{t-1}). \tag{8}$$

通过选取适当的步长，文献[14]给出了平均收敛速率：

$$E[f(\hat{\boldsymbol{w}}_t)-f(\boldsymbol{w}^*)]\leqslant O(1/\sqrt{t}). \tag{9}$$

在光滑目标函数条件下，由于NAG方法进行每一步迭代时都会使用之前迭代的部分甚至全部信息，所以通常可以取得较Heavy-ball方法更快的收敛速率.当目标函数光滑且强凸时，梯度下降方法、Heavy-ball方法与NAG均可以达到线性收敛，即：

$$f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)\leqslant q^t[f(\boldsymbol{w}_0)-f(\boldsymbol{w}^*)], \tag{10}$$

但Heavy-ball方法获得的收敛因子q是最小的[13].

文献[19]通过在投影次梯度上进行了线性插值的操作：

$$\begin{aligned}\boldsymbol{w}_t^+&=\boldsymbol{P}_Q(\boldsymbol{w}_{t-1}^+-\alpha_t\eta_t\nabla f(\boldsymbol{w}_t)),\\ \boldsymbol{w}_{t+1}&=\frac{A_t}{A_{t+1}}\boldsymbol{w}_t+\frac{\alpha_{t+1}}{A_{t+1}}\boldsymbol{w}_t^+,\end{aligned} \tag{11}$$

其中，$Q=\{\boldsymbol{w}:\|\boldsymbol{w}\|_1\leqslant z\}$，$z>0$.该方法在每次迭代时虽然获得了个体解的最优性，但由于在投影操作之后又对所有$\boldsymbol{w}_t$进行了一次加权求和的运算，导致了稀疏性的缺失.文献[20]采用NAG步长策略同样得到了个体收敛的最优性：

$$\begin{aligned}\boldsymbol{y}_t&=\boldsymbol{w}_t+\theta_t(\theta_{t-1}^{-1}-1)(\boldsymbol{w}_t-\boldsymbol{w}_{t-1}),\\ \boldsymbol{w}_{t+1}&=\boldsymbol{P}_Q(\boldsymbol{y}_t-\eta_t\nabla f(\boldsymbol{y}_t)),\end{aligned} \tag{12}$$

其中，θ_t与η_t为步长参数，与线性插值技巧不同的是该方法每一步的解都是通过投影直接得到，因此

可以得到良好的稀疏效果.与之类似,Heavy-ball 方法的个体解也应当具备稀疏性.

2 个体最优收敛性分析

本节给出 Heavy-ball 方法在目标函数非光滑条件下的个体收敛性证明.

对于光滑的优化问题,文献[13]引进加权的动量项 $\boldsymbol{p}_t=\frac{\beta}{1-\beta}(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})$,此时 Heavy-ball 方法的迭代方式可以转化为

$$\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1}=\boldsymbol{w}_t+\boldsymbol{p}_t-\frac{\alpha}{1-\beta}\nabla f(\boldsymbol{w}_t). \quad (13)$$

从式(13)可以看出,Heavy-ball 型动量方法是在梯度下降法基础上添加了动量项 $\boldsymbol{p}_t$.正是由于与梯度下降法的这种相似性,使得梯度下降法的收敛分析思路也可以用于 Heavy-ball 方法.

值得注意的是,文献[13]将 α 和 β 均设定为常数,但对于非光滑优化问题,这样的选取办法无法获得个体收敛速率.因此我们选取了时变的 α 与 β,但此时又会导致式(13)的迭代关系不成立.为了解决这个问题,我们设置 $\boldsymbol{p}_t=t(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})$,通过巧妙地选取 α_t 和 β_t(见定理 1),我们得到:

$$\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1}=\boldsymbol{w}_t+\boldsymbol{p}_t-\frac{2\alpha_t}{1-\beta_t}f(\boldsymbol{w}_t).$$

基于这个关系式,我们可以证明定理 1.为了解决变步长和权重导致的递归问题,我们先证明引理 1.

引理 1. 令 $F=\max\limits_{\boldsymbol{w},\boldsymbol{u}\in Q}d(\boldsymbol{w},\boldsymbol{u})$,$G=\max\limits_{\boldsymbol{w}\in Q}f(\boldsymbol{w})$,有:

$$\sum_{k=1}^{t}\frac{1}{2\eta_k}[(\boldsymbol{w}_k-\boldsymbol{w}^*)^2-(\boldsymbol{w}_{k+1}-\boldsymbol{w}^*)^2]+\frac{\|G\|^2}{2}\sum_{k=1}^{t}\eta_k\leqslant\|F\|^2\frac{1}{2\eta_t}+\frac{\|G\|^2}{2}(2\sqrt{t}-1).$$

具体证明见附录 1.

定理 1. 设 $f(\boldsymbol{w})$ 为一般凸函数,取 $\beta_t=\frac{t}{t+2}$,$\alpha_t=\frac{1}{(t+2)\sqrt{t}}$,$\boldsymbol{w}_t$ 由式(8)产生,则:

$$f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)\leqslant\frac{\sqrt{t}}{2(1+t)}\|F\|^2+\frac{2\sqrt{t}-1}{2(1+t)}\|G\|^2+\frac{1}{1+t}(f(\boldsymbol{w}_0)-f(\boldsymbol{w}^*)).$$

具体证明见附录 2.

综上,我们成功得到了 Heavy-ball 方法在非光滑情况下的最优个体收敛速率.然而批处理形式的 Heavy-ball 方法在计算 $\nabla f(\boldsymbol{w})$ 时需要遍历整个样本集合,这种操作不适合处理大规模数据.为此,我们将上述算法推广至随机形式以求解机器学习问题.

仅考虑二分类问题,假设训练样本集 $S=\{(\boldsymbol{x}_i,y_i)\mid i=1,2,\cdots,m\}\subseteq\mathbb{R}^n\times\{+1,-1\}$,其中 $(\boldsymbol{x}_i,y_i)$ 是独立同分布的.

考虑非光滑稀疏学习问题的损失函数为“hinge 损失”,即 $f_i(\boldsymbol{w})=\max\{0,1-y_i\langle\boldsymbol{w},\boldsymbol{x}_i\rangle\}$ 的优化目标函数为

$$\min_{\boldsymbol{w}\in Q}f(\boldsymbol{w})=\frac{1}{m}\sum_{i=1}^{m}f_i(\boldsymbol{w}). \quad (14)$$

约束情况下随机形式的 Heavy-ball 算法的迭代步骤自然可以表示为

$$\boldsymbol{w}_{t+1}=\boldsymbol{P}_Q(\boldsymbol{w}_t-\alpha_t\nabla f_i(\boldsymbol{w}_t)+\beta_t(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})), \quad (15)$$

其中 i 为迭代到第 t 步时随机抽取的样本序号.

与批处理形式不同的是,随机优化方法的迭代步骤中的 $\nabla f_i(\boldsymbol{w}_t)$ 是 $f_i(\boldsymbol{w}_t)$ 在 t 处的次梯度.由于 hinge 损失函数的次梯度有多种计算方式,这里我们采用文献[26]的方式进行计算,即:

$$f_i(\boldsymbol{w}_t)=\frac{1}{m}\sum_{(\boldsymbol{x}_i,y_i)\in A_t^+}y_i\boldsymbol{x}_i, \quad (16)$$

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

算法 1. 随机 Heavy-ball 算法.

输入:循环次数 t;

输出:$\boldsymbol{w}_t$.

① 初始化向量 $\boldsymbol{w}_1=\boldsymbol{0}$;

② For $k=1$ to t

③ Update $\beta_k=\frac{k}{k+2}$,$\alpha_k=\frac{1}{(k+2)\sqrt{k}}$;

④ 随机选取 $i\in\{1,2,\cdots,m\}$;

⑤ 由式(16)计算 $\nabla f_i(\boldsymbol{w}_k)$;

⑥ 由式(15)计算 $\boldsymbol{w}_{k+1}$;

⑦ End For

当样本点独立同分布时,经过随机抽取样本方式计算得到的 $\nabla f_i(\boldsymbol{w}_t)$ 就是 $f(\boldsymbol{w})$ 在 $\boldsymbol{w}_t$ 处次梯度的无偏估计.从算法 1 中可以看出,随机形式的算法就是将批处理形式的目标函数梯度替换为其无偏估计.文献[29]给出了将批处理算法的收敛界转换为随机算法收敛界的技巧,该技巧对定理 1 同样成立.与文献[19-20,29]完全类似,我们可以将定理 1 推广至随机形式得到定理 2.

定理 2. 设 $f(\boldsymbol{w})$为一般凸函数，取 $\beta_t=\frac{t}{t+2}$，$\alpha_t=\frac{1}{(t+2)\sqrt{t}}$，$\boldsymbol{w}_t$ 由式(15)产生，则：

$$E[f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)]\leqslant\frac{\sqrt{t}}{2(1+t)}\|F\|^2+\frac{2\sqrt{t}-1}{2(1+t)}\|G\|^2+\frac{1}{1+t}(f(\boldsymbol{w}_0)-f(\boldsymbol{w}^*)).$$

根据定理 2，随机 Heavy-ball 方法具有最优的个体收敛速率 $O(1/\sqrt{t})$.

3 实 验

本节对算法 1 的个体收敛速率及其稀疏性的理论分析进行实验验证.

3.1 实验数据集和比较算法

实验所采用的 6 个常用标准数据集，分别为 ijcnn1，covtype，a9a，CCAT，RCV1，astro-physic.数据集来源于 LIBSVM 网站①.表 1 给出了这 6 个数据集的详细描述.

Table 1 Introduction of Standard Datasets

表 1 标准数据集描述

Datasets	Training Samples	Test Samples	Dimensions	Sparsity/%
ijcnn1	49 990	91 701	22	59.09
covtype	522 911	58 101	54	22.12
a9a	24 703	7 858	123	11.27
CCAT	23 149	781 265	47 236	0.16
RCV1	20 242	677 399	47 236	0.16
astro-physic	29 882	32 487	99 757	0.08

实验采用 5 种随机优化方法进行比较，这些方法分别为平均形式输出的标准投影次梯度方法[25,27]、线性插值投影次梯度方法[19]、NAG 方法[20]、平均形式输出的 Heavy-ball 方法[14]以及个体形式输出的 Heavy-ball 方法.从理论分析的角度来说，这 5 种随机优化方法的收敛速率均达到了最优.但在稀疏性方面，个体形式输出的 Heavy-ball 方法与 NAG 方法应该具有较好的表现，而平均形式输出的 Heavy-ball 方法、线性插值投影次梯度方法与标准的投影次梯度方法的稀疏性应该较差.

3.2 实验方法及结论

为公平起见，各算法在每个数据集上均运行 10 次并取平均值作为最后输出.投影次梯度算法及以平均形式输出的 Heavy-ball 算法步长为常数，计算方法分别取自文献[14,27]，即 $\alpha=1/\sqrt{t}$、$\alpha=1/\sqrt{t+1}$，$\beta=0.8$，迭代次数 $t=10\,000$.线性插值投影次梯度算法与 NAG 方法的步长均与迭代次数有关，根据文献[19-20]分别取 $\eta_k=1/\sqrt{k}$ 与 $\alpha_k=1/k\sqrt{k}$.在本实验中，我们调用 SLEP 工具箱的函数来实现投影的计算[28]，其中 $\boldsymbol{P}_Q$ 为 l_1 范数球$\{\boldsymbol{w}:\|\boldsymbol{w}\|_1\leqslant z\}$上的投影算子，根据数据集的不同，$z$ 对应选取不同的值，并且各算法均取相同的约束参数.

图 1 为 5 种算法的收敛速率对比图，其中纵坐标表示当前目标函数值与目标函数最优值之差.粉色实线与蓝色实线分别表示标准的投影次梯度方法与平均形式输出的 Heavy-ball 方法的收敛趋势，青绿色虚线与红色虚线表示线性插值投影次梯度方法与 NAG 方法的收敛趋势，绿色虚线则表示本文提出的以个体形式输出的 Heavy-ball 方法的收敛趋势.从图 1 可以看出，5 种算法在 6 个标准数据集上运行了约 5 000 步之后，基本都达到 10^{-2} 的精度，可以说均表现出基本相同的收敛趋势，这与理论分析的结果是吻合的.

图 2 给出了 5 种算法在 6 个标准数据集上的稀疏性对比，纵坐标表示各算法对应输出方式的稀疏度.稀疏性通过稀疏度来衡量，稀疏度是指变量中非零向量所占的百分比，所以稀疏度越高则稀疏性越差.从图 2 可以看出，线性插值投影次梯度方法虽然以个体形式输出，但稀疏性较差.而 Heavy-ball 方法与 NAG 方法个体解的稀疏度近乎相同，且都明显低于以平均形式输出的投影次梯度方法及 Heavy-ball 方法.由此可知，Heavy-ball 方法的个体输出较好的保留了个体收敛在稀疏性上独具的优势.

另外，从图 2 中还可以看出，对于维数较低的前 3 个数据库，个体解的稀疏性明显优于平均解，基本接近数据集的稀疏度(见表 1 所示)，这充分说明个体解比平均解能更好地描述样本集的稀疏性.但个体解的稀疏度却存在着震荡现象，这主要是由于算法的随机性和稀疏度的分母较小导致的.对维数较高的后 3 个数据集，个体解同样可以描述数据集的稀疏度，但稀疏度已经不再震荡，与平均解一样平稳.

① https://www.csie.ntu.edu.tw/~cjlin/libsvm/

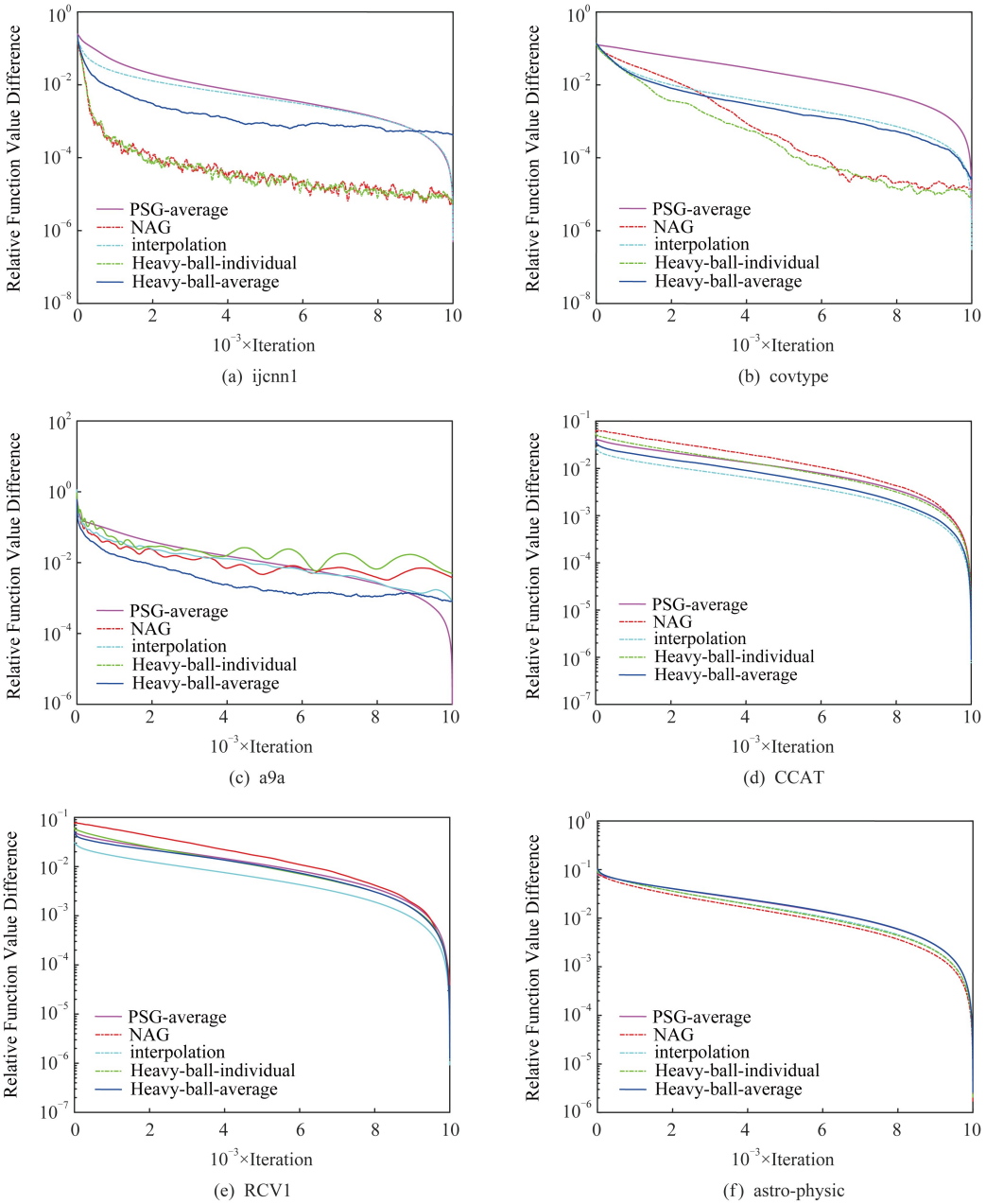

Fig. 1 Comparison of convergence rate

图 1 收敛速率比较图

4 结 论

与其他优化方法相比,Heavy-ball 型动量优化方法目前所知的主要优势是在目标函数强凸且二次可微的条件下获得的收敛速率是最快的.本文对非光滑条件下 Heavy-ball 型动量优化方法的收敛性进行了初步的研究,证明了这种方法可以获得最优的个体收敛速率.众所周知,在不改变算法的情况下,梯度下降方法目前最好的个体收敛速率是 Shamir 和 Zhang 得到的与最优收敛速率差一个 log 因子的个体收敛速率[18].显然,本文的结论表明 Heavy-ball 型动量技巧是对梯度下降法个体收敛速率的一种加速策略,并且与 NAG 方法具有相同的性能.下一步我们将考虑 Heavy-ball 型动量优化方法在正则化和强凸条件下的最优个体收敛速率问题,我们还会考虑在随机 Heavy-ball 型动量优化方法中引进方差减少技巧进一步提升实际收敛效果.

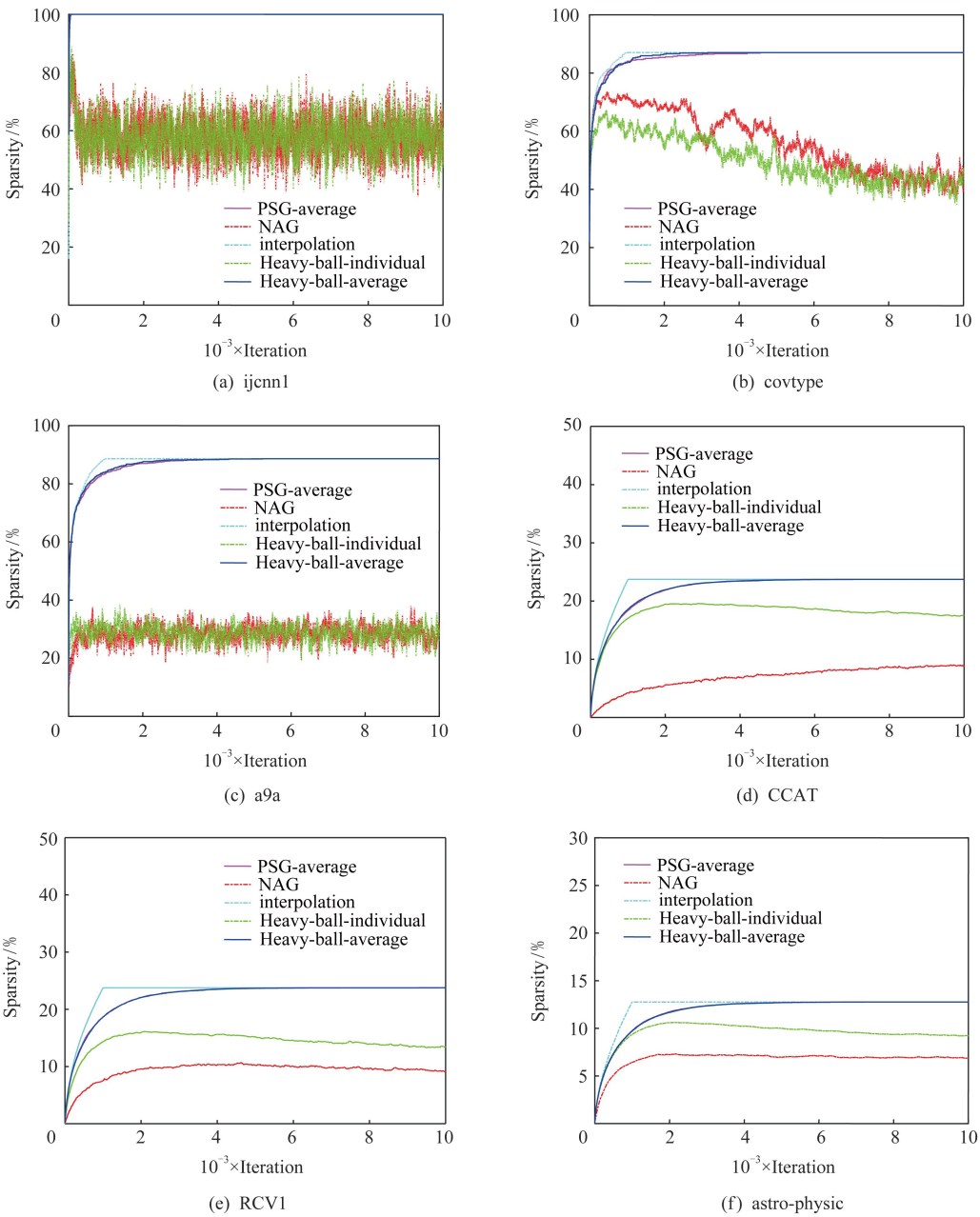

Fig. 2 Comparison of sparsity

图 2 稀疏度比较图

参 考 文 献

[1] Bottou L, Curtis F E, Nocedal J. Optimization methods for large-scale machine learning [J]. SIAM Review, 2018, 60(2): 223-311

[2] 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

[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] Sutskever I, Martens J, Dahl G E, et al. On the importance of initialization and momentum in deep learning [C] //Proc of the Int Conf on Machine Learning (ICML). New York: ACM, 2013: 1139-1147

[5] Nemirovsky A S, Yudin D B. Problem Complexity and Method Efficiency in optimization [M]. New York: Wiley-Interscience, 1983

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

[7] Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems [J]. SIAM Journal on Imaging Sciences, 2009, 2(1): 183-202

[8] Hu C, Kwok J T, Pan W. Accelerated gradient methods for stochastic optimization and online learning [C] //Proc of the 23rd Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2009: 781-789

[9] Lin H, Mairal J, Harchaoui Z. A universal catalyst for first-order optimization [C] //Proc of the 29th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2015: 3384-3392

[10] Allen-Zhu Z. Katyusha: The first direct acceleration of stochastic gradient methods [J]. The Journal of Machine Learning Research, 2017, 18(1): 8194-8244

[11] Mahdavi M, Zhang L, Jin R. Mixed optimization for smooth functions [C] //Proc of the 27th Int Conf on Neural Information Processing Systems. New York: Curran Associates, 2013: 674-682

[12] Shang Fanhua, Jiao Licheng, Zhou Kaiwen, et al. ASVRG: Accelerated proximal SVRG [C] //Proc of the 10th Asian Conf on Machine Learning. Cambridge: JMLR, 2018: 815-830

[13] Ghadimi E, Feyzmahdavian H R, Johansson M. Global convergence of the heavy-ball method for convex optimization [C] //proc of the 14th European Control Conf. Piscataway, NJ: IEEE, 2015: 310-315

[14] Yang Tianbao, Lin Qihang, Li Zhe. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization [EB/OL]. (2016-05-04) [2018-12-13]. https://arxiv.org/abs/1604.03257

[15] Xiao Lin. Dual averaging methods for regularized stochastic learning and online optimization [J]. Journal of Machine Learning Research, 2010, 11(1): 2543-2596

[16] Xiao Lin, Zhang Tong. A proximal stochastic gradient method with progressive variance reduction [J]. SIAM Journal on Optimization, 2014, 24(4): 2057-2075

[17] Shamir O. Open problem: Is averaging needed for strongly convex stochastic gradient descent? [C] //Proc of the 25th Conf on Learning Theory. New York: ACM, 2012: 471-473

[18] Shamir O, Zhang Tong. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes [C] //Proc of the 29th Int Conf on Machine Learning. New York: ACM, 2013: 71-79

[19] Tao Wei, Pan Zhisong, Zhu Xiaohui, et al. The Optimal individual convergence rate for the projected subgradient method with linear interpolation operation [J]. Journal of Computer Research and Development, 2017, 54(3): 529-536 (in Chinese)
(陶蔚, 潘志松, 朱小辉, 等. 线性插值投影次梯度方法的最优个体收敛速率[J]. 计算机研究与发展, 2017, 54(3): 529-536)

[20] Tao Wei, Pan Zhisong, Chu Dejun, 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 (in Chinese)
(陶蔚, 潘志松, 储德军, 等. 使用 Nesterov 步长策略投影次梯度方法的个体收敛性[J]. 计算机学报, 2018, 41(1): 164-176)

[21] Tao Wei, Pan Zhisong, Wu Gaowei, et al. Primal averaging: A new gradient evaluation step to attain the optimal individual convergence [J]. IEEE Transactions on Cybernetics, 2018. DOI: 10.1109/TCYB. 2018. 2874332

[22] Zinkevich M. Online convex programming and generalized infinitesimal gradient ascent [C] //Proc of the 20th Int Conf on Machine Learning. New York: ACM, 2003: 928-936

[23] Duchi J, Shalev-Shwartz S, Singer Y, et al. Efficient projections onto the l_1-ball for learning in high dimensions [C] //Proc of the 25th Int Conf on Machine learning. New York: ACM, 2008: 272-279

[24] Liu Jun, Ye Jieping. Efficient Euclidean projections in linear time [C] //Proc of the 26th Annual Int Conf on Machine Learning. New York: ACM, 2009: 657-664

[25] 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

[26] Shalev-Shwartz S, Singer Y, Srebro N, et al. Pegasos: Primal estimated sub-gradient solver for svm [J]. Mathematical Programming, 2011, 127(1): 3-30

[27] Duchi J C. Introductory lectures on stochastic optimization [EB/OL]. 2010 [2018-12-18]. http://web.standford.edu/~jduchi

[28] Liu Jun, Ji Shuiwang, Ye Jieping. SLEP: Sparse learning with efficient projections [EB/OL]. (2010-10-08) [2018-12-31]. http://www.public.asu.edu/~jye02/Software/SLEP

[29] Rakhlin A, Shamir O, Sridharan K. Making gradient descent optimal for strongly convex stochastic optimization [C] //Proc of the 29th Int Conf on Machine Learning. New York: ACM, 2012: 449-456

Cheng Yujia, born in 1996. Master candidate. Her main research interests include convex optimization algorithms and its application in machine learning.

Tao Wei, born in 1991. PhD candidate. His main research interests include convex optimization algorithms and its application in machine learning.

Liu Yuxiang, born in 1992. Master candidate. His main research interests include convex optimization algorithms and its application in machine learning.

Tao Qing, born in 1965. Professor and PhD supervisor. Senior member of CCF. His main research interests include pattern recognition, machine learning and applied mathematic

附录 1. 正文引理 1 证明.

我们使用 Zinkevich 证明在线优化时采用的迭代技巧[22]进行整理:

$$\sum_{k=1}^{t}\left\{\frac{1}{2\eta_k}[(\boldsymbol{w}_k-\boldsymbol{w}^*)^2-(\boldsymbol{w}_{k+1}-\boldsymbol{w}^*)^2]+\frac{\|G\|^2}{2}\eta_k\right\}\leqslant\frac{1}{2\eta_1}(\boldsymbol{w}_1-\boldsymbol{w}^*)^2-\frac{1}{2\eta_t}(\boldsymbol{w}_{t+1}-\boldsymbol{w}^*)^2+\frac{1}{2}\sum_{k=2}^{t}\left(\frac{1}{\eta_k}-\frac{1}{\eta_{k-1}}\right)(\boldsymbol{w}_k-\boldsymbol{w}^*)^2+\sum_{k=1}^{t}\frac{\|G\|^2}{2}\eta_k\leqslant\|F\|^2\left[\frac{1}{2\eta_1}+\frac{1}{2}\sum_{k=1}^{t}\left(\frac{1}{\eta_k}-\frac{1}{\eta_{k-1}}\right)\right]+\frac{\|G\|^2}{2}\sum_{k=1}^{t}\eta_k\leqslant\|F\|^2\frac{1}{2\eta_t}+\frac{\|G\|^2}{2}\sum_{k=1}^{t}\eta_k.$$

当 $\eta_t=\frac{1}{\sqrt{t}}$时,

$$\sum_{k=1}^{t}\eta_k=\sum_{k=1}^{t}\frac{1}{\sqrt{k}}\leqslant1+\int_{k=1}^{t}\frac{\mathrm{d}k}{\sqrt{k}}\leqslant1+[2\sqrt{k}]_1^t\leqslant2\sqrt{t}-1.$$

引理 1 得证.

附录 2. 正文定理 1 证明.

根据式(13)有:

$$\|\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1}-\boldsymbol{w}^*\|^2\leqslant\|\boldsymbol{w}_t+\boldsymbol{p}_t-\boldsymbol{w}^*\|^2-\frac{4\alpha_t}{1-\beta_t}\langle\boldsymbol{w}_t+t(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})-\boldsymbol{w}^*,\nabla f(\boldsymbol{w}_t)\rangle+\left(\frac{2\alpha_t}{1-\beta_t}\right)^2\|f(\boldsymbol{w}_t)\|_2^2=\|\boldsymbol{w}_t+\boldsymbol{p}_t-\boldsymbol{w}^*\|^2-\frac{4\alpha_t}{1-\beta_t}\langle\boldsymbol{w}_t-\boldsymbol{w}^*,\nabla f(\boldsymbol{w}_t)\rangle-\frac{4\alpha_t t}{1-\beta_t}\langle\boldsymbol{w}_t-\boldsymbol{w}_{t-1},\nabla f(\boldsymbol{w}_t)\rangle+\left(\frac{2\alpha_t}{1-\beta_t}\right)^2\|\nabla f(\boldsymbol{w}_t)\|_2^2.$$

将 $\beta_t=\frac{t}{t+2}$,$\alpha_t=\frac{1}{(t+2)\sqrt{t}}$代入:

$$\|\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1}-\boldsymbol{w}^*\|^2\leqslant\|\boldsymbol{w}_t+\boldsymbol{p}_t-\boldsymbol{w}^*\|^2-\frac{2}{\sqrt{t}}\langle\boldsymbol{w}_t-\boldsymbol{w}^*,\nabla f(\boldsymbol{w}_t)\rangle-\frac{2t}{\sqrt{t}}\langle\boldsymbol{w}_t-\boldsymbol{w}_{t-1},\nabla f(\boldsymbol{w}_t)\rangle+\frac{1}{t}\|f(\boldsymbol{w}_t)\|_2^2.$$

不等式左右两边同乘$\sqrt{t}$:

$$\sqrt{t}\|\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1}-\boldsymbol{w}^*\|^2+2(t+1)(f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*))\leqslant\sqrt{t}\|\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1}-\boldsymbol{w}^*\|^2+2t(f(\boldsymbol{w}_{t-1})-f(\boldsymbol{w}^*))+\frac{1}{\sqrt{t}}\|\nabla f(\boldsymbol{w}_t)\|_2^2,$$

$$2(t+1)(f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)\leqslant2(f(\boldsymbol{w}_0)-f(\boldsymbol{w}^*))+\sum_{k=1}^{t}\sqrt{k}(\|\boldsymbol{w}_t+\boldsymbol{p}_t-\boldsymbol{w}^*\|^2-\|\boldsymbol{w}_{t+1}+\boldsymbol{p}_{t+1}-\boldsymbol{w}^*\|^2)+\sum_{k=1}^{t}\frac{1}{\sqrt{k}}\|\nabla f(\boldsymbol{w}_k)\|_2^2.$$

根据引理 1 得:

$$f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)\leqslant\frac{\sqrt{t}}{2(1+t)}\|F\|^2+\frac{2\sqrt{t}-1}{2(1+t)}\|G\|^2+\frac{1}{1+t}(f(\boldsymbol{w}_0)-f(\boldsymbol{w}^*)).$$

定理 1 得证.