使用 Nesterov 步长策略投影次梯度方法的个体收敛性

陶 蔚[1)] 潘志松[1)] 储德军[2)] 陶 卿[2)]

[1)](中国人民解放军陆军工程大学指挥信息系统学院 南京 210007)

[2)](中国人民解放军陆军炮兵防空兵学院十一系 合肥 230031)

摘 要 很多机器学习问题都可以最终转换为优化问题来进行求解,凸优化算法已经被成功用于各种机器学习优化问题中,而在优化算法的研究中是否能获得最优的收敛速率是一个最基本问题.此外,稀疏性是稀疏学习问题中关注的另一个目标.目前,人们已经提出了大量的随机优化方法求解大规模机器学习优化问题,但大部分的研究只是针对平均输出方式获得了最优收敛速率.个体输出方式显然比平均方式的输出具有更好的稀疏性,但使个体收敛速率获得最优具有一定的难度,人们已经将强凸情形下的最优个体收敛性作为公开问题进行广泛研究.对于光滑目标函数的优化问题,著名学者 Nesterov 提出了一种步长策略,使得梯度方法的收敛速率获得了数量级形式的加速,并且获得了最优的个体收敛速率.目前,Nesterov 加速算法已经应用于各种具有光滑损失函数机器学习优化问题中,研究者基于该加速策略提出了大量的随机优化算法.能否将这种技巧推广至非光滑情形获得最优的个体收敛速率显然是有意义的问题.文中考虑在非光滑优化算法中引入这种步长策略.特别地,我们聚焦经典的一阶梯度方法,提出了一种嵌入加速算法步长策略的投影次梯度算法,证明了这种算法在求解非光滑损失函数学习问题时具有最优的个体收敛速率.这是比标准投影次梯度方法只有在平均输出方式下才具有最优收敛速率更强的结论,也是一阶梯度方法在个体最优收敛速率方面比较接近于大家期待的研究成果.与平均方式输出以及线性插值的投影次梯度方法相比,该文所提方法的梯度运算在插值策略之后,因此在求解 $l1$ 范数约束的 hinge 损失函数学习问题时具有更好的稀疏性.人工数据集上的实验验证了所提方法的正确性,基准数据集上验证了该方法在保持稀疏性方面具有良好的性能.

关键词 机器学习;非光滑损失函数问题;投影次梯度方法;Nesterov 步长策略;个体收敛速率;稀疏学习

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2018.00164

The Individual Convergence of Projected Subgradient Methods Using the Nesterov's Step-Size Strategy

TAO Wei[1)] PAN Zhi-Song[1)] CHU De-Jun[2)] TAO Qing[2)]

[1)](*College of Command Information System, Army Engineering University of PLA, Nanjing* 210007)

[2)](*11st Department, PLA Army Academy of Artillery and Air Defense, Hefei* 230031)

Abstract Many problems arising in machine learning can be finally reduced to optimization problems. Convex optimization algorithms have been successfully adapted in various kinds of learning optimization problems. And whether the optimal convergence rate can be attained is one of the basic problems in the study of optimization algorithms. Besides, sparsity is another concern in sparse learning problems. So far, a great deal of stochastic optimization algorithms have been presented for solving the large scale learning problems. However, most of the-state-of-the-arts stochastic optimization algorithms only attain the optimal convergence rates in terms of the averaged output, and the desired sparsity can not be guaranteed. In contrast to the averaged

收稿日期:2016-12-10;在线出版日期:2017-04-11.本课题得到国家自然科学基金(61273296,61673394)资助.**陶 蔚**,男,1991 年生,博士研究生,主要研究方向为凸优化算法及其在机器学习中的应用、网络安全.E-mail:wtao_plaust@163.com.**潘志松**,男,1973 年生,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为模式识别、机器学习和网络安全.**储德军**,男,1978 年生,博士研究生,讲师,主要研究方向为模式识别、机器学习.**陶 卿**(通信作者),男,1965 年生,博士,教授,博士生导师,中国计算机学会(CCF)高级会员,主要研究领域为模式识别、机器学习和应用数学.E-mail:qing.tao@ia.ac.cn.

output, the individual solution usually offers more sufficient sparsity. Unfortunately, it is not easy to make the individual convergence rate optimal and the optimal individual convergence rate in strongly-convex cases has been extensively exploring as an open problem. For solving smooth objective optimization problems, it is well known that the step-size rule raised by the famous researcher Nesterov's can accelerate the convergence rate of the first order gradient algorithm by orders of magnitude, and the optimal individual convergence rate are simultaneously derived. Recently, Nesterov's acceleration algorithm has been commonly applied in various learning optimization problem with smooth loss functions, and a large number of stochastic optimization algorithms in smooth cases have been developed based on the Nesterov's acceleration strategy. Obviously, whether the Nesterov's step-size rule can be extended to obtain the optimal individual convergence rate for nonsmooth objective optimization problems is an interesting problem. In this paper, the Nesterov's step-size rule in smooth objective cases is incorporated into the gradient method for solving nonsmooth objective optimization problems. In particular, focusing on the classic first order gradient methods, we present a new projected subgradient method with the Nesterov's step-size rule. It is proved that the proposed method can achieve the optimal individual convergence rate when solving nonsmooth optimization problems. Such conclusion is stronger than the previous one that the regular projected subgradient method can obtain the optimal convergence result only in terms of the averaged output. And it can also be regarded as an approximate answer to the question of whether first order gradient methods can achieve the optimal individual convergence rate. Compared with the regular projected subgradient methods in which the averaged output is used or the modified projected subgradient methods in which the linear interpolation operation is employed, the subgradient-like operation follows the extrapolation evaluation in our method, which brings significant benefits in keeping the sufficient sparsity when solving the hinge loss function optimization problems on an $l1$-norm ball. The experiments on two synthetic datasets verify that our theoretical analysis is correct, and the experiments on several benchmark datasets demonstrate that the proposed methods have almost the same convergence behavior but offer more sufficient sparsity. As future work, the optimal individual convergence in regularized sparse learning problems and the stability of individual convergence in stochastic optimization will be considered. Moreover, by using the Nesterov's step-size rule, whether the optimal individual convergence for strongly-convex objective functions can be achieved will be investigated.

Keywords　machine learning; nonsmooth optimization problems; projected subgradient method; Nesterov's step-size rule; individual convergence rate; sparse learning

1　引　言

正则化经验风险最小化是机器学习监督学习算法设计中普遍遵循的一个准则[1]，其中经验风险是每个训练样本导致的损失之和. 在实际应用中，不同的分类问题往往采用不同的损失函数. 常用的光滑损失主要有最小二乘法和对数函数，而非光滑损失主要指的是 hinge 函数[1]. 由于光滑性的不同，研究者们在进行求解时，使用的一阶梯度方法无论在步长策略方面还是在最优收敛速率方面，均存在着很大的差异.

目前，光滑目标函数的优化问题成果丰硕，吸引了众多研究者的关注. 显然，由于函数光滑性的提高，我们可以选取更加快速收敛的常数步长，同时使得标准的梯度方法比非光滑情形的收敛速率有着数量级式的提高. 然而，更吸引研究者的工作应该是 Nesterov 于 1983 年提出的一种开创性的加速方法[2]，它可以将多种形式一阶梯度方法的收敛速率提升一个数量级，并且所获得的收敛速率是最优的.

具体来说,Nesterov 加速算法的核心思想是使用一种步长技巧,与标准梯度方法不同的是其每一步的迭代会使用之前迭代的部分甚至全部信息[3-4].由于随机优化方法在处理学习问题时,具有计算代价低和实际收敛速度快等优点,已经成为处理大规模问题的主流方法.目前,对于正则化损失函数但仅需正则化项凸和损失函数光滑的优化问题,一些随机优化方法也采用了对算法加速起关键作用的步长策略,成功获得了随机情形下的最优收敛速率[5-6].

一个有趣的问题是,如果在非光滑问题的优化算法中使用光滑问题加速算法的步长策略,会得到怎样的结果?由于投影次梯度方法是梯度下降法对非光滑约束优化问题的一种推广[7],且几乎所有的非光滑优化方法都在此基础上发展而来.另外从计算角度来说,$l1$ 范数球上投影高效计算方法的出现使得投影次梯度成为求解稀疏学习问题的一种具有竞争力的方法[8-9],因此本文主要考虑在投影次梯度算法中引入这种步长策略,得到了最优个体收敛速率这一有意义而且意想不到的结果.

相比于光滑损失函数,非光滑损失函数优化方法的研究显得较为平淡,这主要是因为基于普通的投影次梯度方法,就可以直接获得一般凸优化问题的最优收敛速率.尽管最优收敛速率已经无法进一步提升,但其中仍然有一些问题,特别是在个体收敛速率方面,没有取得令人满意的结果.

机器学习随机优化算法收敛速率的早期研究几乎完全依赖于其对应在线优化算法 regret 界的分析[10-11].利用凸函数的基本性质,可以很容易通过凸优化问题在线算法的 regret 界得到随机算法的收敛速率[12],这其中包括 Pegasos[13]、RDA(Regularized Dual Average)[6] 和 COMID(Composite Objective MIrror Descent)[14].虽然通过这种方法获得的收敛速率仅仅是针对所有迭代进行平均的输出方式,但由于可以获得一般凸问题的最优收敛速率,人们还是可以接收这种平均方式的收敛性.但遗憾的是,有例子表明即使是采用平均形式,SGD(Stochastic Gradient Descent)在求解强凸目标函数时仍然达不到最优的收敛速率界[15],由此引发了一些学者对平均输出方式甚至是经典 SGD 优化方法本身的一些质疑.另一方面,由于在求解稀疏学习问题时平均方式输出结果的稀疏性通常比个体解要差很多,个体收敛速率又重新回到人们的视野,成为更加值得关注的目标.但个体最优收敛速率的研究需要面临很多实际困难.早在 2012 年,强凸目标函数下 SGD 的个体最优收敛速率问题就已经正式被列为 open 问题[16].

事实上,单纯地讨论个体收敛性并不存在很大的障碍.特别地,从 SGD 已有的平均收敛速率结论就可以很容易直接获得其个体的收敛速率,但由于比平均方式的收敛界多出了一个对数因子,所获收敛速率界的最优性无法达到[17].为了进一步获得个体收敛的最优性,人们采用了适当修改已有随机优化算法的思路,对偶平均优化算法成为人们首先关注的目标.例如,Chen 等人[18]采用了在迭代过程中增加一种梯度运算步骤的方式,而 Nesterov 和 Shikhman[19]则增添了一种线性插值操作.文献[18]的方法对凸性和光滑性不同的优化问题均获得了最优的个体收敛速率,该方法也因此被称为最优的对偶平均优化方法,而文献[19]的方法只是针对凸问题得到了最优的个体收敛速率.但相比而言,最优的对偶平均优化方法由于增添了优化子问题的求解,对标准方法的改动比较大,而线性插值技巧与标准的对偶平均方法区别极小,可以认为是一种良好扩展.需要指出的是,这两种改进方法都只关注了对偶平均优化方法,并没有讨论其它形式的一阶梯度优化方法.

在文献[20]中,我们成功地将线性插值技巧拓展至投影次梯度方法,得到了最优个体收敛速率.但是,稀疏性方面的表现却与平均形式输出解相差无几,这主要是由于最终输出解由线性插值方式产生,无法获得充分的稀疏性.本文提出一种嵌入光滑优化问题加速算法步长策略的投影次梯度算法,证明了这种算法在求解非光滑损失函数学习问题时具有最优的个体收敛速率.这是比标准投影次梯度方法只有在平均输出方式下才具有最优收敛速率更强的结论,并且这种嵌入步长策略方法与标准算法改动不大,比较接近大家的期待.由于步长策略嵌入在迭代过程的梯度运算之前,与平均方式输出以及线性插值的投影次梯度方法相比,所提方法在求解 $l1$ 范数约束的 hinge 损失函数学习问题时具有更好的稀疏性.实验也验证了所提方法在处理稀疏学习问题时理论分析的正确性以及在保持稀疏性方面良好的性能.

2 光滑损失加速算法的步长策略

Nesterov 提出的加速步长策略主要针对黑箱形式的光滑目标函数优化问题,其主要思路是将优化变量每一次迭代与上一次迭代适当组合在一起进

行梯度运算，进一步通过对步长进行精心选择的初始化和迭代，从而获得加速效果[2]. 目前，这种加速方法被扩展到正则化只需损失函数光滑的情形，并且其步长策略也出现了多种形式的变形[3-4]. 注意到加速步长策略不同变形下的收敛性证明会使用不同的技巧[4]，本节仅围绕正则化光滑损失函数问题对所涉及的加速步长策略做简单必要的介绍[4].

考虑下列无约束的优化问题：

$$\min \widetilde{F}(\boldsymbol{w})=\lambda\|\boldsymbol{w}\|_1+\widetilde{f}(\boldsymbol{w}) \tag{1}$$

因为 $\widetilde{f}(\boldsymbol{w})$是光滑的，即存在，$L>0$ 满足：

$$\widetilde{f}(\boldsymbol{w})+\widetilde{f}(\boldsymbol{v})\leq\langle\partial\widetilde{f}(\boldsymbol{v}),\boldsymbol{w}-\boldsymbol{v}\rangle+\frac{L}{2}\|\boldsymbol{w}-\boldsymbol{v}\|^2,$$

对于具有稀疏正则化项的问题如式(1)，近邻梯度(Proximal Gradient，PG)方法在迭代过程中保持正则化项不动，仅使用损失函数的二次展开进行运算[3-4]，即

$$\boldsymbol{w}_{t+1}=\arg\min_{\boldsymbol{w}}\left\{\lambda\|\boldsymbol{w}\|_1+\widetilde{f}(\boldsymbol{w}_t)+\langle\partial\widetilde{f}(\boldsymbol{v}),\boldsymbol{w}-\boldsymbol{w}_t\rangle+\frac{L}{2}\|\boldsymbol{w}-\boldsymbol{w}_t\|^2\right\} \tag{2}$$

该方法有如下收敛速率界：

$$\widetilde{F}(\boldsymbol{w}_t)-\widetilde{F}(\boldsymbol{w})\leq O\left(\frac{L}{t}\right) \tag{3}$$

Tseng 在文献[4]中系统介绍了三种处理光滑损失函数优化问题的加速方法，人们统称为加速近邻梯度(Accelerated Proximal Gradient，APG)方法. 本文仅涉及其中的第一种加速方法，这种加速方法与文献[3]中的 FISTA(Fast Iterative Shrinkage Thresholding Algorithm)算法也是一样的，即

$$\begin{cases}\boldsymbol{y}_t=\boldsymbol{w}_t+\theta_t(\theta_{t-1}^{-1}-1)(\boldsymbol{w}_t-\boldsymbol{w}_{t-1}),\\ \boldsymbol{w}_{t+1}=\arg\min_{\boldsymbol{w}}\left\{\lambda\|\boldsymbol{w}\|_1+f(\boldsymbol{y}_t)+\langle\nabla f(\boldsymbol{y}_t),\boldsymbol{w}-\boldsymbol{y}_t\rangle+\frac{L}{2}\|\boldsymbol{w}-\boldsymbol{y}_t\|^2\right\},\\ \theta_{t+1}=(\sqrt{\theta_t^4+4\theta_t^2}-\theta_t^2)/2\end{cases} \tag{4}$$

其中 $t\geq 1$.

与式(2)简单形式迭代不同的是，式(4)中使用了一种插值形式的步长策略，在本文中我们将这种形式的步长方法称为 Nesterov 步长策略. 可以证明，APG 的收敛速率界为

$$\widetilde{F}(\boldsymbol{w}_t)-\widetilde{F}(\boldsymbol{w})\leq O\left(\frac{L}{t^2}\right) \tag{5}$$

显然，在光滑问题的梯度算法中引入了 Nesterov 步长策略后，收敛速率获得了最优并且比普通梯度算法的收敛速率有了数量级形式的提升.

目前，APG 算法已经被成功推广至随机情形，对于光滑损失函数的正则化优化问题获得了关于迭代步数以及方差形式的最优收敛速率[5-6]. 在这些随机优化方法中，也普遍采用了批处理方法中的 Nesterov 步长策略. 本文主要考虑在求解非光滑优化问题的投影次梯度方法中引入 Nesterov 步长策略，尽管无法做到收敛速率数量级形式的提升，但却可以获得最优的个体收敛速率.

3　稀疏学习问题与投影次梯度方法

本文仅考虑二分类问题. 假设训练样本集合 $S=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}\in\mathbb{R}^n\times\{+1,-1\}$，其中$(\boldsymbol{x}_i,y_i)$之间满足独立同分布. 并且我们只考虑最简单但最典型的非光滑稀疏学习问题“$l1$ 正则化＋hinge 损失”，即

$$\min_{\boldsymbol{w}}\lambda\|\boldsymbol{w}\|_1+\frac{1}{m}\sum_{i=1}^{m}f_i(\boldsymbol{w}) \tag{6}$$

这里 $f_i(\boldsymbol{w})=\max\{0,1-y_i(\boldsymbol{w},\boldsymbol{x}_i)\}$，$\lambda$ 是平衡参数，反映了稀疏性和分类损失之间的一种折中.

对于给定的平衡参数 λ，存在 $z>0$，使得问题(6)具有下列等价的优化形式：

$$\min_{\boldsymbol{w}}f(\boldsymbol{w})=\frac{1}{m}\sum_{i=1}^{m}f_i(\boldsymbol{w}),\quad \text{s.t. } \|\boldsymbol{w}\|_1\leq z \tag{7}$$

对于式(7)，批处理形式投影次梯度方法的迭代步骤为

$$\boldsymbol{w}_{t+1}=\boldsymbol{P}(\boldsymbol{w}_t-\alpha_t\nabla f(\boldsymbol{w}_t)) \tag{8}$$

其中 $\boldsymbol{P}(\cdot)$为 $l1$ 范数球 $\{\boldsymbol{w}:\|\boldsymbol{w}\|_1\leq z\}$ 上的投影算子，α_t是迭代步长，$\nabla f(\boldsymbol{w}_t)$是 $f(\boldsymbol{w})$在 $\boldsymbol{w}_t$处的次梯度. 该算法随机形式的迭代步骤为

$$\boldsymbol{w}_{t+1}=\boldsymbol{P}(\boldsymbol{w}_t-\alpha_t\nabla f(\boldsymbol{w}_t)) \tag{9}$$

其中，$\nabla f_i(\boldsymbol{w}_t)$是 $f_i(\boldsymbol{w})$在 $\boldsymbol{w}_t$处的次梯度，i 是第 t 次迭代随机抽取的样本序号.

通过将批处理算法的目标函数梯度换成其无偏估计，可以得到随机算法平均解的收敛速率界：

定理 1[21]. 记 $\boldsymbol{w}^*=\arg\min_{\|\boldsymbol{w}\|_1\leq z}f(\boldsymbol{w})$. 假设 $\forall\boldsymbol{w}\in\{\boldsymbol{w}:\|\boldsymbol{w}\|_1\leq z\}$和 $i\in\{1,\cdots,m\}$，都有 $\|\nabla f_i(\boldsymbol{w})\|_1\leq C$ 成立. 任取 $\boldsymbol{w}_0\in\mathbb{R}^n$，设 $\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_T$由式(9)产生，则当步长 $\alpha_t=t^{-1/2}$时，

$$E[f(\bar{\boldsymbol{w}}_T)-f(\boldsymbol{w}^*)]\leq O(1/\sqrt{T}),$$

其中 $\bar{\boldsymbol{w}}_T=(\boldsymbol{w}_1+\cdots+\boldsymbol{w}_T)/T$.

由定理 1 可知，对于一般凸的优化问题，平均输出方式的随机投影次梯度方法可以得到最优的收敛

结果,该算法的个体收敛界是 $O(\ln T/\sqrt{T})$[17],但是尚未能达到最优.

受 Nesterov[19] 在对偶平均方法中嵌入线性插值操作得到一般凸情形下最优个体收敛速率的启发,我们在文献[20]中提出了一种嵌入线性插值操作的投影次梯度方法:

$$\begin{aligned}&\boldsymbol{w}_t^+=\boldsymbol{P}(\boldsymbol{w}_{t-1}^+-\alpha_t\eta_t\nabla f(\boldsymbol{w}_t)),\\&\boldsymbol{w}_{t+1}=\frac{A_t}{A_{t+1}}\boldsymbol{w}_t+\frac{\alpha_{t+1}}{A_{t+1}}\boldsymbol{w}_t^+\end{aligned}\tag{10}$$

其中,$t\geq 1$,$A_k=\sum_{t=0}^{k}\alpha_t$,$\alpha_t$ 和 η_t 均为步长参数,算法中的第 t 步输出是 $\boldsymbol{w}_t$. 值得注意的是,文献[19]直接在对偶平均方法中嵌入线性插值操作就可得到最优个体收敛速率,但这种做法对投影次梯度方法来说却是行不通的. 式(10)与标准投影次梯度方法的主要区别是,该算法沿 $\boldsymbol{w}_{t-1}^+$ 方向做投影梯度运算,而不是投影次梯度方法中的 $\boldsymbol{w}_t$ 方向. 针对一般凸目标优化问题,该算法可以得到 $O(1/\sqrt{T})$ 的最优的个体收敛结果.

注意到式(10)中,算法的最终输出 $\boldsymbol{w}_{t+1}$ 由插值得到,当将这个插值的式子展开时,$\boldsymbol{w}_{t+1}$ 可以表示之前迭代产生的所有 $\boldsymbol{w}_t^+$ 的组合形式. 由此可知,在处理稀疏学习问题时,$\boldsymbol{w}_{t+1}$ 的稀疏程度与平均输出方式的解并无本质的差别,本文的实验也验证了这一点. 实际上,如何改进投影次梯度方法使其具有个体最优收敛速率,并在处理稀疏学习问题中确保个体解的良好稀疏性,是本文的主要动机.

4 引入步长策略的投影次梯度方法

本节我们在非光滑优化问题中引入 Nesterov 加速步长策略,可以得到一种具有最优个体收敛速率的投影次梯度方法,并将所获结论推广至随机方法情形.

对于问题(7),具体算法如下:

$$\begin{aligned}&\boldsymbol{y}_t=\boldsymbol{w}_t+\theta_t(\theta_{t-1}^{-1}-1)(\boldsymbol{w}_t-\boldsymbol{w}_{t-1}),\\&\boldsymbol{w}_{t+1}=\boldsymbol{P}(\boldsymbol{y}_t-\eta_t\nabla f(\boldsymbol{y}_t))\end{aligned}\tag{11}$$

其中 $t\geq 1$,步长策略满足:$(1-\theta_{t+1})/\theta_{t+1}^2\leq 1/\theta_t^2$,且 $\theta_t\in(0,1]$. 在上述条件下,一种通常的取法是 $\theta_{t+1}=(\sqrt{\theta_t^4+4\theta_t^2}-\theta_t^2)/2$,另一种取法为 $\theta_t=2/(t+2)$[4].

从形式上看,式(11)和(10)似乎区别不大,只是采用了不同的步长策略而已. 但式(11)未对投影次梯度方法进行任何改变,因此一旦证明了个体最优收敛速率,与线性插值的方法相比,可以说该方法是对于个体收敛速率问题更加接近大家期待的研究成果. 另一个重要区别是 Nesterov 加速步长在梯度投影运算之前,进而对于稀疏学习问题,投影运算有效保证了个体解的稀疏性. 而在式(10)中,插值操作在梯度投影运算之后,其稀疏性无法得到保证.

下面对加入步长策略的投影次梯度方法(式(11))进行收敛性分析.

引理 1[22]. 假设其中 $\boldsymbol{P}_Q$ 是闭凸集合 Q 上的投影算子,则对任意 $\boldsymbol{w}\in\mathbb{R}^n$,$\boldsymbol{w}_0\in Q$,则

$$\langle\boldsymbol{w}-\boldsymbol{w}_0,\boldsymbol{u}-\boldsymbol{w}_0\rangle\leq 0,$$

对任意 $\boldsymbol{u}\in Q$ 都成立的充要条件是:

$$\boldsymbol{w}_0=\boldsymbol{P}_Q(\boldsymbol{w}).$$

引理 2. 对于任意 $\boldsymbol{y}_t\in Q$,

$$\eta_t(f(\boldsymbol{y}_t)-f(\boldsymbol{y}))\leq\langle\eta_t\nabla f(\boldsymbol{y}_t),\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\rangle+\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{y}_t\|^2-\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{w}_{t+1}\|^2-\frac{1}{2}\|\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\|^2.$$

引理 3. $\eta_t(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{y}))\leq\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{y}_t\|^2-\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{w}_{t+1}\|^2+\frac{\eta_t^2}{2}M^2$,其中 M 是 hinge 损失函数的 Lipschitz 常数.

定理 2. 假设 $\eta_t\leq\eta_{t-1}$,$f(\boldsymbol{w})$ 是闭凸集合 $Q\in\mathbb{R}^n$ 上的凸函数,投影区域 Q 为有界集合,即 $\|\boldsymbol{x}-\boldsymbol{y}\|\leq M_1$,$\forall\boldsymbol{x},\boldsymbol{y}\in Q$,有

$$f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w})\leq\theta_t^2\left(\frac{1}{2\eta_{t+1}}M_1^2+\sum_{k=0}^{t+1}\frac{\eta_k}{2\theta_k^2}M^2\right).$$

根据定理 2,我们可得出以下推论:

推论. $f(\boldsymbol{w})$ 是有界闭凸集合 $Q\in\mathbb{R}^n$ 上的凸函数. 取 $\theta_t=\frac{2}{t+2}$ 和 $\eta_t=\frac{1}{(t+2)\sqrt{t+2}}$,对于任意 $\boldsymbol{w}\in Q$,则有

$$f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w})\leq O\left(\frac{M_1^2+M^2}{\sqrt{t}}\right).$$

我们进一步将式(11)推广至随机情形,即

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

其中 $t\geq 1$,且 $\nabla f_i(\boldsymbol{y}_t)$ 是 $f(\boldsymbol{y}_t)$ 在 $\boldsymbol{y}_t$ 处的次梯度.

在标准的机器学习问题中,人们往往假定样本集合里的样本是独立同分布的,因此每次随机抽取得到的 $\nabla f_i(\boldsymbol{y}_t)$ 是 $\nabla f(\boldsymbol{y}_t)$ 的一个无偏估计. 此时,随机优化算法的主要操作就是把批处理算法使用的目标函数梯度换成其无偏估计. 文献[15]的引理 1 给出了一种将批处理算法的收敛速率转化成随机优化算法收敛速率的证明技巧,这种技巧对于非光滑问

题是成立的. 类似地，我们可以获得期望条件下引理 2 和引理 3 对应的结论，并进一步将定理 2 的结论推广至随机情形，即得到定理 3.

定理 3.　假设 $\eta_t \leq \eta_{t-1}$，$f(\boldsymbol{w})$是闭凸集合 $Q \in \mathbb{R}^n$ 上的凸函数，投影区域 Q 为有界集合，即 $\|\boldsymbol{x}-\boldsymbol{y}\| \leq M_1$，$\forall \boldsymbol{x}, \boldsymbol{y} \in Q$，则有

$$E(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w})) \leq \theta_t^2\left(\frac{1}{2\eta_{t+1}}M_1^2+\sum_{k=0}^{t+1}\frac{\eta_k}{2\theta_k^2}M^2\right).$$

在定理 3 的基础上，很容易将上述推论也推广至随机情形，即 Nesterov 加速步长策略的随机投影次梯度方法具有个体最优收敛速率. 由于标准随机投影次梯度方法从平均方式解的收敛速率可以直接推导出个体解的收敛速率，但未能达到最优[17]. 因此，定理 3 是比投影次梯度方法平均方式最优收敛速率更强的结论.

众所周知，hinge 损失是机器学习中最常用也是最典型的非光滑损失函数，为了说明个体收敛性在保持机器学习问题稀疏性方面的良好性能，本文的理论分析和实验部分主要考虑求解 $l1$ 范数约束的 hinge 损失函数优化问题. 实际上，本文所获得的结论只假定了目标函数的凸性及次梯度的有界性（见定理 2 和定理 3），并不依赖于 hinge 损失函数，完全适用于其它非光滑凸损失函数，具体应用中只是表现为次梯度不同而已.

5　实　验

本节对引入 Nesterov 步长策略的随机投影次梯度方法在处理稀疏学习问题时个体收敛速率、解的正确性和稀疏性进行实验验证.

所有的实验采取随机方法抽取样本，各算法均取相同的约束参数和步长，每个算法分别运行 10 次，实验的最终输出取这 10 次运行结果的平均值.

由于 hinge 损失函数的次梯度不是唯一的，可以有多种方式进行计算，这里我们采取和文献[13]完全相同方式进行计算，即

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

其中 $A_t \subseteq S$，$A_t^+=\{(\boldsymbol{x}_i,y_i)\in A_t : y_i\langle \boldsymbol{w},\boldsymbol{x}_i\rangle<1\}$，且 $|A_t|=k$. 实验中我们设置 k 为 1，即每次迭代仅随机选取 1 个训练样本进行次梯度的计算.

5.1　人工数据集上的实验

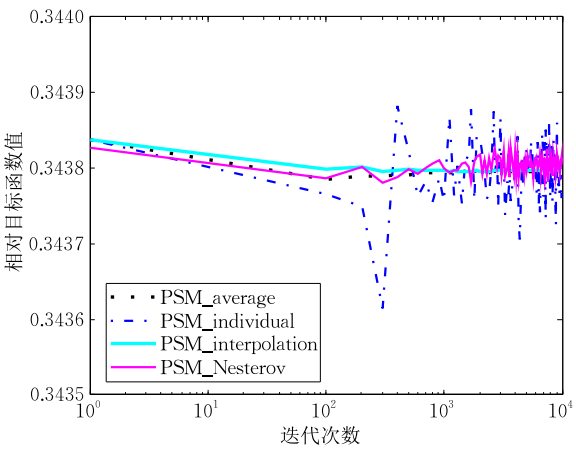
为了对所提出的加入 Nesterov 步长策略随机投影次梯度方法所得求解结果的正确性进行评估，首先给出 2 个人工数据集（synthetic data1 和 synthetic data2）的基本描述.

Duchi 等人在文献[8]中构造了人工数据集，很好地验证了随机投影次梯度方法中投影算子算法的正确性. 本节人工数据集的构造原理与文献[8]相同，即首先根据期望为 0 的正态分布产生一个向量 $\boldsymbol{w}_0$，并随机等可能性地令其中 20% 的分量为 0 元素；接着以同样的随机方式产生 10 000×10 000 的训练样本. 每个样本的标签 y_i 由 $\boldsymbol{w}_0$ 与 $\boldsymbol{x}_i$ 乘积的符号决定，并随机等可能性改变其中 10% 的标签. 实验中我们构造了 2 个数据集，表 1 给出了这 2 个数据集的基本描述.

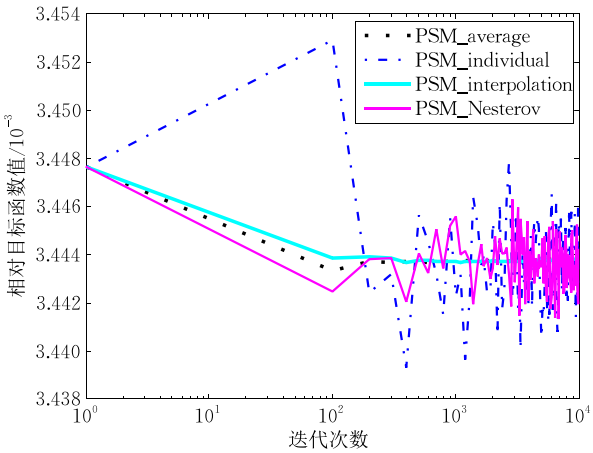
表 1　人工数据集的基本描述

数据集	训练样本数	维数	方差
synthetic data1	10 000	10 000	0.01
synthetic data2	10 000	10 000	0.0001

图 1 给出了几种算法的收敛趋势. 实验结果比



(a) synthetic data1

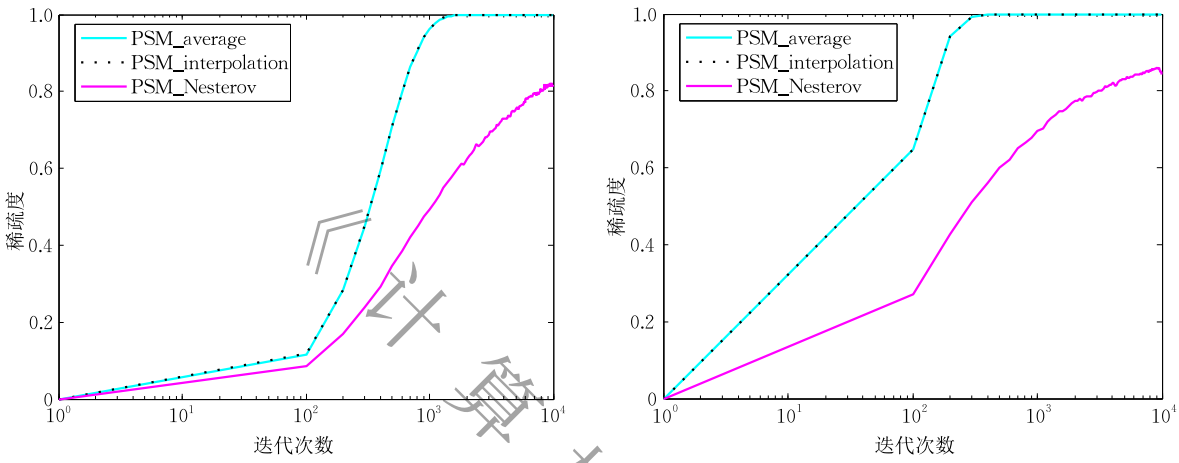


(b) synthetic data2

图 1　人工数据集收敛速率比较图

较图采用半对数坐标系，横坐标取对数坐标，表示迭代次数，最大迭代次数设置为 10000 次，纵坐标取直角坐标，表示相对目标函数值，即当前目标函数值与最优目标函数值之差的值. 这里最优目标函数值是通过初始设置的 w_0 计算得到. 图 1 中的黑色短虚线(PSM_average)描述了投影次梯度方法平均输出方式解的收敛趋势，蓝色点虚线(PSM_individual)描述了投影次梯度方法个体输出方式的收敛趋势，天蓝色实线(PSM_interpolation)表示文献[19]嵌入线性插值投影次梯度方法个体解的收敛趋势，而粉色曲线(PSM_Nesterov)则表示本文提出的引入 Nesterov 步长策略投影次梯度方法个体解的收敛趋势.

接下来在人工数据集上进行稀疏性的比较，实验中采用了稀疏度这个指标来衡量稀疏性的好坏. 稀疏度表示一个向量中非零分量所占的百分比，稀疏度越高零元素越少，因而稀疏性就越差[23]. 在图 2 中，横坐标仍然是迭代次数，纵坐标表示 10 次实验稀疏度的平均值.



(a) synthetic data1　　(b) synthetic data2

图 2　人工数据集稀疏度比较图

根据图 1 的收敛结果可知，对于 synthetic data1，迭代到一定步骤以后，3 种不同形式的投影次梯度方法均很快收敛达到了 $f(\boldsymbol{w}_t)-f(\boldsymbol{w}_0)\leq 4\times 10^{-1}$ 的精度；对于另一个数据集 synthetic data2，收敛曲线则可以达到精度为 $f(\boldsymbol{w}_t)-f(\boldsymbol{w}_0)\leq 10^{-2}$，这在一定程度上验证了所提算法收敛性分析的正确性.

从图 2 可以看出，在两个不同的人工数据集上，本文方法所得解均能接近初始 w_0 设置的稀疏度，并且优于平均输出方式和嵌入线性插值策略的投影次梯度方法，这就说明了所提算法在保持稀疏性方面的良好性能.

另外，与 synthetic data1 相比，所提方法在 synthetic data2 上具有更接近 $f(\boldsymbol{w}_0)$ 的精度，这也反映了实验数据噪声误差对于实验结果的影响.

5.2　基准数据集上的实验

在本节实验中，我们的目的是验证所提算法在基准数据集上的表现. 实验选择了 6 个基准数据集，即 a9a、w8a、covtype、RCV1、CCAT 和 astro-physics，这些基准数据集来源于 LIBSVM① 网站. 表 2 给出了它们的基本描述.

表 2　基准数据集的基本描述

数据集	训练样本数	维数	稀疏度/%
a9a	24703	123	11.27
w8a	49749	300	3.88
covtype	522911	54	22.12
RCV1	20242	47236	0.16
CCAT	23149	47236	0.16
astro-physic	29882	99757	0.08

6 个基准数据集收敛速率的比较结果如图 3 所示，其中横坐标、纵坐标以及各曲线的含义与图 1 相同. 由于基准数据集目标函数的最优值未知，因此我们取各算法迭代过程中 10 次平均后最小的目标函数值作为最优值.

从图 3 可以看出，本文提出的引入 Nesterov 步长策略投影次梯度方法的个体解与投影次梯度方法平均输出解以及嵌入线性插值操作投影次梯度方法的个体解均具有相同的收敛趋势，这就从实验上验证了本文所提方法具有最优的个体收敛速率. 尽管投影次梯度方法的个体输出方式(PSM_individual)也具有相同的收敛趋势，但却无此方面的收敛速

① www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

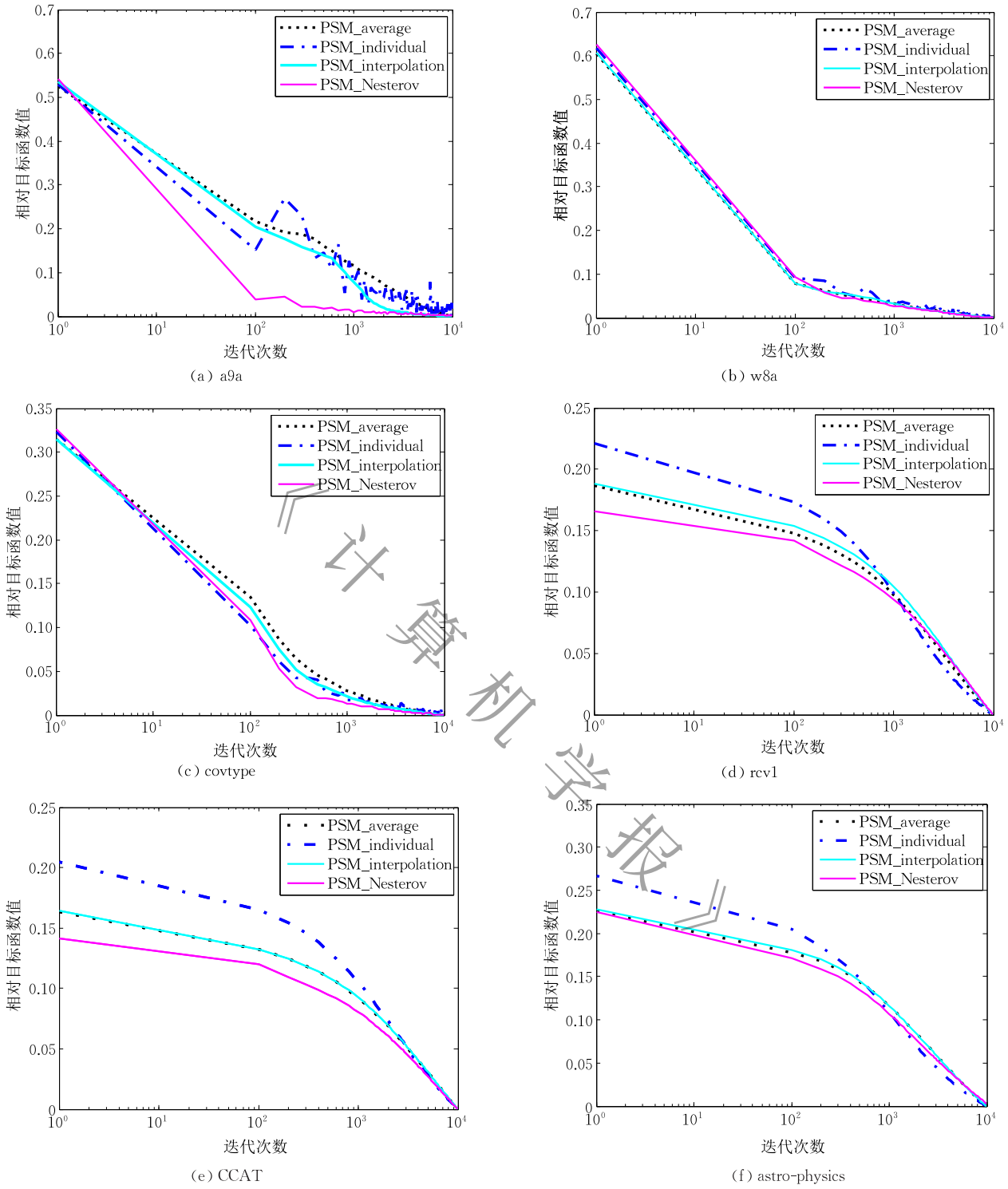
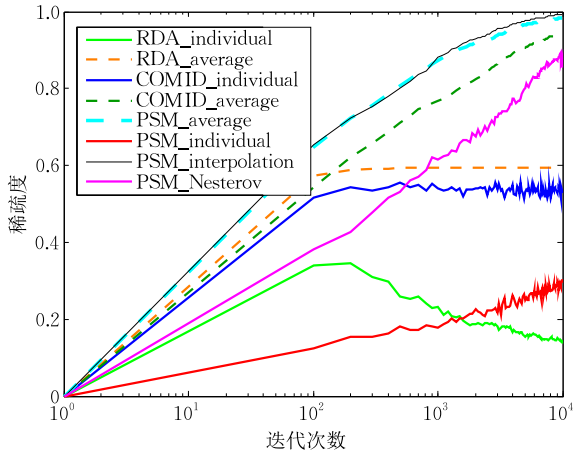


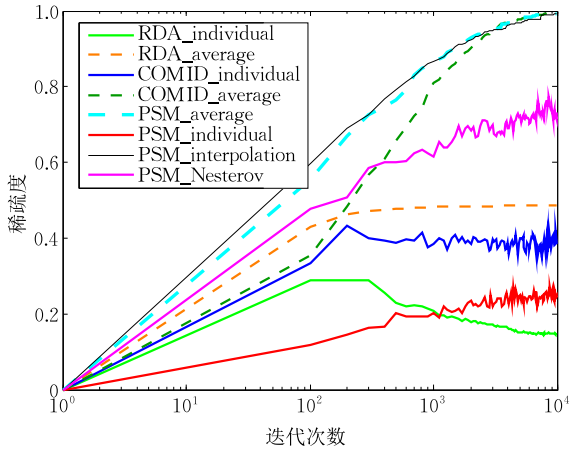
图 3　基准数据集收敛速率比较图

率界.

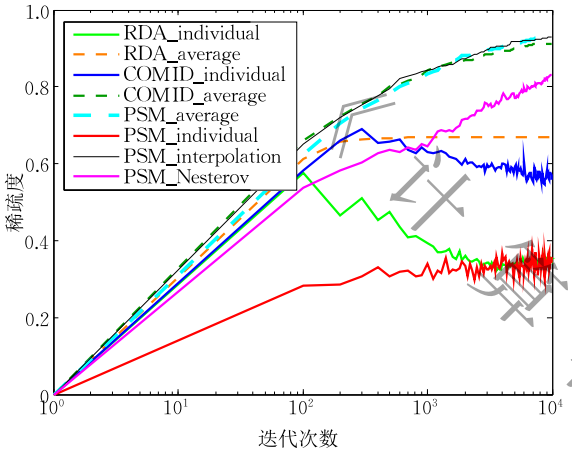
在稀疏度的比较方面，我们将比较对象进行了扩充，几乎包括了目前主流的求解非光滑优化问题的一阶随机梯度方法，实验比较结果如图 4 所示. 横坐标和纵坐标与图 2 相同. 为了增强对比效果，本文不仅给出了投影次梯度各种输出方式结果的稀疏度，也显示了 RDA 和 COMID 算法两种输出方式的稀疏结果. 其中实线表示的是各种算法个体解的稀疏性结果，分别是投影次梯度方法(PSM_individual)、线性插值投影次梯度(PSM_interpolation)、RDA(RDA_individual)和 COMID 方法(COMID_individual). 虚线均表示平均方式输出解稀疏性的变化趋势，分别是投影次梯度(PSM_average)、RDA(RDA_average)和 COMID 方法(COMID_average).



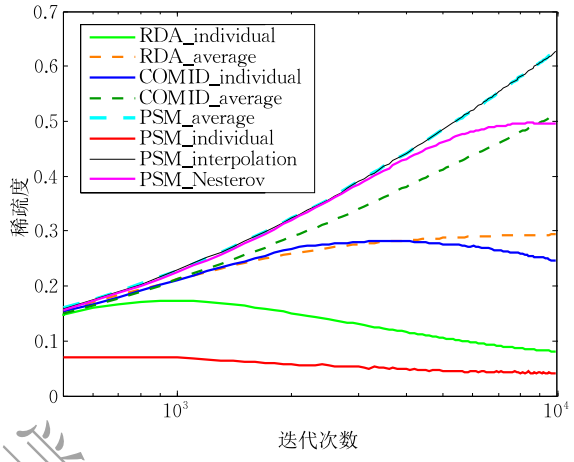
(a) a9a



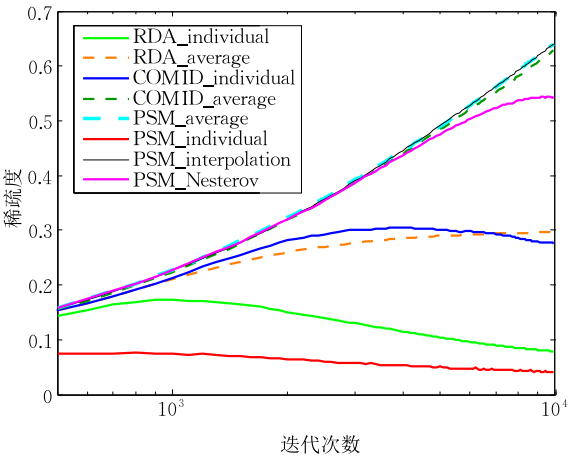
(b) w8a



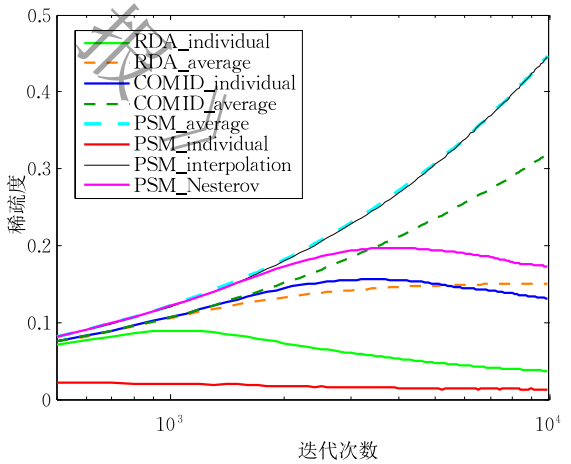
(c) covtype



(d) rcv1



(e) CCAT



(f) astro-physics

图 4 基准数据集稀疏度比较图

其中,粉色的曲线(PSM_Nesterov)是本文所提方法在迭代过程中稀疏度的变化曲线.

从图 4 可以看出,不管是投影次梯度,RDA 还是 COMID 方法,个体解的稀疏性都要明显比平均方式解的稀疏性要好. 对于平均方式的解,RDA 和 COMID 的稀疏性要比 PSM 方法稀疏性要好,这体现了 RDA 和 COMID 即使是平均解也能够较好地保证 $l1$ 范数稀疏性的特点. 对于个体解,本文所提方法的稀疏性要明显好于线性插值形式的投影次梯度方法,这从实验上验证了引入 Nesterov 步长策略的投影次梯度方法在求解稀疏优化问题中保持稀疏性方面的良好性能.

从图 4 中还可以看出，本文算法个体解的稀疏性比 RDA 和 COMID 个体解的稀疏性要差，但 RDA 和 COMID 的个体解是否具有最优的个体收敛速率目前仍然不得而知，这也是 Shamir 的 open 问题所关注的内容. 显然，本文所提方法的优势在于具有最优个体收敛速率的理论保证，并且在实验中个体解的稀疏性要比投影次梯度方法平均输出方式以及线性插值方式投影次梯度方法个体解的稀疏性均有一定程度的改善.

需要指出的是，本文所提方法在选择不同参数情形下会出现不稳定的收敛趋势，稀疏性也会随之发生明显震荡. 正如文献[21]所指出的那样，这些现象是随机优化算法个体解在收敛过程中不可避免的缺陷，是仅由个体计算目标函数值的随机性产生的，取迭代过程后期一定数目权值的平均值计算目标函数就会明显改善这种不稳定性，这和 α-suffix 平均[15]的主要思路是一致的. 需要指出的是，我们提出的嵌入线性插值的算法也获得了最优的个体收敛速率，但是插值操作在算法操作过程中起到了加权平均的作用，从而导致了算法个体收敛具有较好的稳定性，但所获得解的稀疏性却表现一般，仅和标准投影次梯度方法平均方式解的稀疏性相当. 本文提出的使用 Nesterov 步长策略投影次梯度方法在收敛性和稀疏性方面的表现和嵌入线性插值的投影次梯度算法正好相反. 显然，能否对随机投影次梯度梯度方法进行适当的改进，获得最优个体收敛速率并高效保证正则化项结构，同时具有收敛稳定性是一个值得研究的问题.

6　总　结

非光滑优化问题最优收敛速率的研究已经有了很多研究成果，但所获得的结论都是平均形式的. 特别对于具有 $l1$ 正则化的稀疏学习问题，虽然每一步输出的权向量 $\boldsymbol{w}$ 都有令人满意的稀疏度，但是最终的平均结果往往差强人意.

众所周知，对于光滑的优化问题，Nesterov 的加速技巧可以将梯度方法的收敛速率提高一个数量级. 本文的主要贡献是发现了这种技巧在非光滑优化问题中的作用，即在非光滑问题中引进 Nesterov 加速算法的步长策略后，可以得到一种具有最优个体收敛速率的投影次梯度方法. 作为应用，这种方法在处理稀疏学习问题时可以充分保证最终解的稀疏性. 研究具有最优个体收敛速率又有收敛稳定性的稀疏随机投影次梯度算法是我们下一步主要的研究方向.

致　谢　本文审稿专家和编辑老师提出了宝贵的意见和建议，在此表示感谢！

参　考　文　献

[1] Zhang Tong. Statistical behavior and consistency of classification methods based on convex risk minimization. Annals of Statistic, 2004, 32(1): 56-85

[2] 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

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

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

[5] Hu C, Kwok J T, Pan W. Accelerated gradient methods for stochastic optimization and online learning//Proceedings of the Conference on Neural Information Processing Systems, Vancouver, Canada. 2009: 781-789

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

[7] Shor N Z. Minimization Methods for Non-Differentiable Functions. Berlin, Germany: Springer, 1985

[8] Duchi J, Shalev-Shwartz S, Singer Y. Efficient projections onto the $l1$-ball for learning in high dimensions//Proceedings of the International Conference on Machine Learning. Helsinki, Finland. 2008: 272-279

[9] Liu Jun, Ye Jie-Ping. Efficient Euclidean projections in linear time//Proceedings of the International Conference on Machine Learning, Montreal, Canada, 2009: 657-664

[10] Zinkevich M. Online convex programming and generalized infinitesimal gradient ascent//Proceedings of the International Conference on Machine Learning. Washington, USA, 2003: 928-936

[11] Hazan E, Agarwal A, Kale S. Logarithmic regret algorithms for online convex optimization. Machine Learning, 2007, 69(2): 169-192

[12] Shalev-Shwartz S. Online learning and online convex optimization. Foundations and Trends in Machine Learning, 2011, 4(2): 107-194

[13] Shalev-Shwartz S, Singer Y, Srebro N. Pegasos: Primal estimated sub-gradient solver for SVM//Proceedings of the International Conference on Machine Learning. Oregon, USA, 2007: 807-814

[14] Duchi J, Shalev-Shwartz S, Singer Y, et al. Composite objective mirror descent//Proceedings of the Conference on Learning Theory. Haifa, Israel, 2010: 116-128

[15] Rakhlin A, Shamir O, Sridharan K. Making gradient descent optimal for strongly convex stochastic optimization//Proceedings of the 29th International Conference on Machine Learning. Edinburgh, Scotland, 2012: 449-456

[16] Shamir O. Open problem: Is averaging needed for strongly convex stochastic gradient descent?//Proceedings of the Conference on Learning Theory. Edinburgh, Scotland, 2012: 47.1-47.3

[17] Shamir O, Zhang Tong. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA, 2013: 71-79

[18] Chen Xi, Lin Qi-Hang, Pena J. Optimal regularized dual averaging methods for stochastic optimization//Proceedings of the Advances in Neural Information Processing Systems. Nevada, USA, 2012: 404-412

[19] Nesterov Y, Shikhman V. Quasi-monotone subgradient methods for nonsmooth convex minimization. Journal of Optimization Theory and Applications, 2015, 165(3): 917-940

[20] Tao Wei, Pan Zhi-Song, Zhu Xiao-Hui, et al. The Optimal individual convergence rate for the projected subgradient method with linear interpolation operation. 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. 计算机研究与发展, 2017, 54(3): 529-536)

[21] Hazan E, Kale S. Beyond the regret minimization barrier: An optimal algorithm for stochastic strongly-convex optimization. Journal of Machine Learning Research, 2014, 15(1): 2489-2512

[22] Bertsekas D P, Nediać A, Ozdaglar A E. Convex Analysis and Optimization. Belmont, USA: Athena Scientific, 2003

[23] Tao Qing, Gao Qian-Kun, Jiang Ji-Yuan, et al. Survey of solving the optimization problems for sparse learning. Journal of Software, 2013, 24(11): 2498-2507(in Chinese)
(陶卿, 高乾坤, 姜纪远等. 稀疏学习优化问题的求解综述. 软件学报, 2013, 24(11): 2498-2507)

附录 1.

引理 2 证明.

$$\begin{aligned}\|\boldsymbol{y}-\boldsymbol{y}_t\|^2&=\|\boldsymbol{y}-\boldsymbol{w}_{t+1}+\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\|^2\\&=\|\boldsymbol{y}-\boldsymbol{w}_{t+1}\|^2+\|\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\|^2+\\&\quad 2\langle\boldsymbol{w}_{t+1}-\boldsymbol{y}_t,\boldsymbol{y}-\boldsymbol{w}_{t+1}\rangle\langle\boldsymbol{w}_{t+1}-\boldsymbol{y}_t,\boldsymbol{y}-\boldsymbol{w}_{t+1}\rangle\\&=\langle\boldsymbol{w}_{t+1}-(\boldsymbol{y}_t-\eta_t\nabla f(\boldsymbol{y}_t)),\boldsymbol{y}-\boldsymbol{w}_{t+1}\rangle-\\&\quad\langle\eta_t\nabla f(\boldsymbol{y}_t),\boldsymbol{y}-\boldsymbol{w}_{t+1}\rangle\\&=\langle\boldsymbol{w}_{t+1}-(\boldsymbol{y}_t-\eta_t\nabla f(\boldsymbol{y}_t)),\boldsymbol{y}-\boldsymbol{w}_{t+1}\rangle-\\&\quad\langle\eta_t\nabla f(\boldsymbol{y}_t),\boldsymbol{y}-\boldsymbol{y}_t+\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\rangle\\&=\langle\boldsymbol{w}_{t+1}-(\boldsymbol{y}_t-\eta_t\nabla f(\boldsymbol{y}_t)),\boldsymbol{y}-\boldsymbol{w}_{t+1}\rangle-\\&\quad\eta_t\langle\nabla f(\boldsymbol{y}_t),\boldsymbol{y}-\boldsymbol{y}_t\rangle-\eta_t\langle\nabla f(\boldsymbol{y}_t),\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\rangle,\end{aligned}$$

根据引理 1, $\langle\boldsymbol{w}_{t+1}-(\boldsymbol{y}_t-\eta_t\nabla f(\boldsymbol{y}_t)),\boldsymbol{y}-\boldsymbol{w}_{t+1}\rangle\geq 0$,

$$\begin{aligned}\langle\eta_t\nabla f(\boldsymbol{y}_t),\boldsymbol{y}_t-\boldsymbol{y}\rangle&\leq\langle\boldsymbol{w}_{t+1}-\boldsymbol{y}_t,\boldsymbol{y}-\boldsymbol{w}_{t+1}\rangle+\\&\quad\eta_t\langle\nabla f(\boldsymbol{y}_t),\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\rangle\\&\leq\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{y}_t\|^2-\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{w}_{t+1}\|^2-\\&\quad\frac{1}{2}\|\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\|^2+\eta_t\langle\nabla f(\boldsymbol{y}_t),\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\rangle.\end{aligned}$$

注意到, $\eta_t\langle\nabla f(\boldsymbol{y}_t),\boldsymbol{y}-\boldsymbol{y}_t\rangle\leq\eta_t\langle\nabla f(\boldsymbol{y}_t),\boldsymbol{y}_t-\boldsymbol{y}\rangle$.

引理 2 得证!

附录 2.

引理 3 证明.

由于 f 满足 Lipschiz 条件, 根据文献[18]中的式(2)可知,

$$f(\boldsymbol{w}_{t+1})\leq f(\boldsymbol{y}_t)+\langle\nabla f(\boldsymbol{y}_t),\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\rangle+M\|\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\|.$$

下面建立 $f(\boldsymbol{w}_{t+1})$ 与 $f(\boldsymbol{y})$ 的联系,

$$\begin{aligned}\eta_t(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{y}))&=\eta_t(f(\boldsymbol{y}_t)-f(\boldsymbol{y}))+\eta_t(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{y}_t))\\&\leq\eta_t\langle\nabla f(\boldsymbol{y}_t),\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\rangle+\\&\quad\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{y}_t\|^2-\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{w}_{t+1}\|^2-\\&\quad\frac{1}{2}\|\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\|^2+\eta_t\langle\nabla f(\boldsymbol{y}_t),\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\rangle+\\&\quad M\eta_t\|\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\|,\end{aligned}$$

注意到, $-\frac{1}{2}\|\boldsymbol{y}_t-\boldsymbol{w}_{t+1}\|^2+M\eta_t\|\boldsymbol{w}_{t+1}-\boldsymbol{y}_t\|\leq\frac{\eta_t^2}{2}M^2$.

引理 3 得证!

附录 3.

定理 2 证明.

证明思路与文献[4]中 Theorem 1(b)相同, 令:

$$\boldsymbol{y}=(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w},$$

由引理 3 可得,

$$\begin{aligned}&\eta_t f(\boldsymbol{w}_{t+1})-\eta_t f((1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w})\\&\leq\frac{1}{2}\|(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w}-\boldsymbol{y}_t\|^2-\\&\quad\frac{1}{2}\|(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w}-\boldsymbol{w}_{t+1}\|^2+\frac{\eta_t^2}{2}M^2,\end{aligned}$$

$$\frac{1}{2}\|(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w}-\boldsymbol{y}_t\|^2-\frac{1}{2}\|(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w}-\boldsymbol{w}_{t+1}\|^2$$

$$=\frac{\theta_t^2}{2}\|\boldsymbol{w}+(\theta_t^{-1}-1)\boldsymbol{w}_t-\theta_t^{-1}\boldsymbol{y}_t\|^2-\frac{\theta_t^2}{2}\|\boldsymbol{w}+(\theta_t^{-1}-1)\boldsymbol{w}_t-\theta_t^{-1}\boldsymbol{w}_{t+1}\|^2,$$

令 $\boldsymbol{z}_t=-(\theta_t^{-1}-1)\boldsymbol{w}_t+\theta_t^{-1}\boldsymbol{y}_t$，

$$\frac{1}{2}\|(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w}-\boldsymbol{y}_t\|^2-\frac{1}{2}\|(1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w}-\boldsymbol{w}_{t+1}\|^2$$

$$=\frac{\theta_t^2}{2}\|\boldsymbol{w}-\boldsymbol{z}_t\|^2-\frac{\theta_t^2}{2}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|^2,$$

$$\eta_t f(\boldsymbol{w}_{t+1})-\eta_t f((1-\theta_t)\boldsymbol{w}_t+\theta_t\boldsymbol{w})\leq\frac{\theta_t^2}{2}\|\boldsymbol{w}-\boldsymbol{z}_t\|^2-\frac{\theta_t^2}{2}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|^2+\frac{\eta_t^2}{2}M^2,$$

即

$$\eta_t(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w}))\leq\eta_t(1-\theta_t)(f(\boldsymbol{w}_t)-f(\boldsymbol{w}))+\frac{\theta_t^2}{2}\|\boldsymbol{w}-\boldsymbol{z}_t\|^2-\frac{\theta_t^2}{2}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|^2+\frac{\eta_t^2}{2}M^2,$$

$$\frac{f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w})}{\theta_t^2}\leq\frac{(1-\theta_t)(f(\boldsymbol{w}_t)-f(\boldsymbol{w}))}{\theta_t^2}+\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_t\|^2-\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|^2+\frac{\eta_t}{2\theta_t^2}M^2,$$

即

$$\frac{(1-\theta_{t+1})(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w}))}{\theta_{t+1}^2}\leq\frac{(1-\theta_t)(f(\boldsymbol{w}_t)-f(\boldsymbol{w}))}{\theta_t^2}+\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_t\|^2-\frac{1}{2\eta_t}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|^2+\frac{\eta_t}{2\theta_t^2}M^2,$$

依次类推，

$$\frac{(1-\theta_{t+1})(f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w}))}{\theta_{t+1}^2}\leq\frac{(1-\theta_0)(f(\boldsymbol{w}_0)-f(\boldsymbol{w}))}{\theta_0^2}+\sum_{k=0}^{t}\left(\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_k\|^2-\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_{k+1}\|^2\right)+\sum_{k=0}^{t}\frac{\eta_k}{2\theta_k^2}M^2$$

$$=\sum_{k=0}^{t}\left(\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_k\|^2-\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_{k+1}\|^2\right)+\sum_{k=0}^{t}\frac{\eta_k}{2\theta_k^2}M^2.$$

由于 $\eta_t\leq\eta_{t-1}$，

$$\sum_{k=1}^{t}\left(\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_k\|^2-\frac{1}{2\eta_k}\|\boldsymbol{w}-\boldsymbol{z}_{k+1}\|^2\right)\leq\frac{1}{2\eta_0}\|\boldsymbol{w}-\boldsymbol{z}_0\|^2-\frac{1}{2\eta_{t+1}}\|\boldsymbol{w}-\boldsymbol{z}_{t+1}\|^2+\sum_{k=0}^{t+1}\left(\frac{1}{2\eta_k}-\frac{1}{2\eta_{k-1}}\right)\|\boldsymbol{w}-\boldsymbol{z}_k\|^2\leq\frac{1}{2\eta_{t+1}}M_1^2.$$

定理 2 证毕！

附录 4.

推论证明.

根据定理 2，

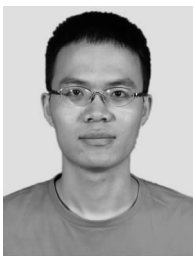
$$f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w})\leq\theta_t^2\left(\frac{1}{2\eta_{t+1}}M_1^2+\sum_{k=0}^{t+1}\frac{\eta_k}{2\theta_k^2}M^2\right),$$

取 $\theta_t=\frac{2}{t+2}$ 和 $\eta_t=\frac{2}{(t+2)\sqrt{t+2}}$，由于

$$\frac{1}{2\eta_{t+1}}\theta_t^2M_1^2\leq O(1/\sqrt{t})\text{ 且 }\theta_t^2\sum_{k=0}^{t+1}\frac{\eta_k}{2\theta_k^2}M^2\leq O(1/t\sqrt{t}),$$

所以，$f(\boldsymbol{w}_{t+1})-f(\boldsymbol{w})\leq O(1/\sqrt{t})$.

推论得证！



TAO Wei, born in 1991, Ph. D. candidate. His main research interests include convex optimization algorithm and its application in machine learning, network security.

PAN Zhi-Song, born in 1973, Ph. D., professor, Ph. D. supervisor. His main research interests include pattern recognition, machine learning and network security.

CHU De-Jun, born in 1978, Ph. D. candidate. His main research interests include pattern recognition, and machine learning.

TAO Qing, born in 1965, Ph. D., professor, Ph. D. supervisor. His main research interests include pattern recognition, machine learning and applied mathematics.

Background

Many problems arising in statistical machine learning can be cast as constrained convex optimization problems, and quite a few interesting adaptations of fundamental optimization algorithms that exploit the structure and fit the requirements of the application have been presented. Due to the explosion in size of modern machine learning problems, stochastic optimization algorithm has been becoming one of the state-of-the-art methods for solving large-scale learning optimization problems. Stochastic optimization scheme offers simple iteration schemes and has shown good practical performance, making it particularly efficient for large-scale learning problems. However, one of the most important problems in theoretical

analysis of stochastic optimization is whether or not the optimal convergence rate can be derived and the learning structure can be ensured.

So far, various kinds of stochastic optimization algorithms have been presented for solving the regularized loss problems. For solving smooth objective optimization problems, the step-size rule raised by the famous researcher Nesterov can accelerate the convergence rate of the gradient-like algorithm by orders of magnitude.

In contrast to smooth cases, there is less exciting work in nonsmooth optimization. This may be because almost all the subgradient-like algorithms can simply derive the optimal convergence rate. While the popular projected subgradient method has been extensively studied in theory and application, there are still several basic problems far from satisfactory. Specifically, the optimal convergence rate in nonsmooth cases is obtained only in terms of averaging of all past iterates due to employment of the standard online-to-batch conversion, and even the simplest sparsity cannot be preserved. In contrast to the averaged output, the individual solution can keep the sparsity very well, and its optimal convergence rate in strongly-convex cases is extensively explored even as an open problem.

In this paper, we focus on the nonsmooth loss learning problems, in which Nesterov's step-size rule is incorporated into the gradient method for solving nonsmooth objective optimization problems. In particular, we present a projected subgradient method with the Nesterov's step-size rule. It is proved that the proposed method can achieve the optimal individual convergence rate when solving nonsmooth optimization problems. Such conclusion is stronger than the previous one that the regular projected subgradient method can obtain the optimal convergence result only in terms of the averaged output. And it can also be regarded as an approximate answer to the question of whether first-order gradient methods can achieve the optimal individual convergence rate. Typically, our method is applicable as an efficient tool for solving large-scale hinge loss optimization problems on an $l1$-norm ball. Compared with the projected subgradient methods in which the averaged output is used or the linear interpolation operation is employed, the proposed method can ensure sufficient sparsity when solving the hinge loss function optimization problems on an $l1$-norm ball. The experiments on synthetic datasets verify the correctness of the proposed method, and the experiments on several benchmark datasets demonstrate its high performance in preserving sparsity. As future work, the optimal individual convergence in regularized sparse learning problems and the stability of individual convergence in stochastic optimization will be considered. Moreover, by using the Nesterov's accelerated step-size rule, whether the optimal individual convergence for strongly-convex objective functions can be achieved will be investigated.

This work is partially supported by the National Natural Science Foundation of China (Nos. 61273296 and 61673394).