随机 COMID 的瞬时收敛速率分析

姜纪远,陶 卿,邵言剑,汪群山
(中国人民解放军陆军军官学院十一系,安徽合肥 230031)

摘 要: COMID(Composite Objective MIrror Descent)是一种能够保证 L1 正则化结构的在线算法,其随机收敛速率可由在线算法的 regret 界直接得到,但其最终解是 T 次迭代平均的形式,稀疏性很差.瞬时解具有很好的稀疏性,因此分析算法的瞬时收敛速率在随机学习中变得越来越重要.本文讨论正则化非光滑损失的随机优化问题,当正则化项为 L1 和 L1 + L2 时,分别证明了 COMID 的瞬时收敛速率.大规模数据库上的实验表明,在保证几乎相同正确率的同时,瞬时解一致地提高了稀疏性,尤其是对稀疏性较差的数据库,稀疏度甚至能够提升 4 倍以上.

关键词: 机器学习; 随机优化; 非光滑优化; L1 正则化; COMID; 瞬时收敛速率
中图分类号: TP301 **文献标识码:** A **文章编号:** 0372-2112 (2015)09-1850-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.09.025

The Analysis of Convergence Rate of Individual COMID Iterates

JIANG Ji-yuan, TAO Qing, SHAO Yan-jian, WANG Qun-shan
(11*th Department*, *Army Officer Academy of PLA*, *Hefei*, *Anhui* 230031, *China*)

Abstract: COMID is an online algorithm which can ensure the structure of L1 regularization. Its stochastic convergence rate can be obtained directly from the regret bound in online settings. However, the derived final solution has poor sparsity because it only takes the form of averaging all previous T iterates. Naturally, the individual solution has nice sparisity. So it becomes more and more important to discuss individual convergence rates in the stochastic learning. In this paper, we focus on the regularized non-smooth loss problems. When the regularizer are L1 and L1 + L2, we prove the individual convergence rates of COMID respectively. The extensive experiments on large-scale datasets demonstrate that the individual solution consistently improves the sparsity while keeping almost the same accuracy. For the datasets with poor sparse structure, the sparsity of solution is improved even up to four times.

Key words: machine learning; stochastic optimization; non-smooth optimization; L1 regularization; COMID; individual convergence rate

1 引言

目前,“正则化 + 损失函数”的函数结构被普遍地作为机器学习问题的优化目标[1].正则化主要用于调节分类器的泛化能力,防止过拟合,常见的正则化有 L1[2]、L2[3]以及 L1 + L2 混合正则化[4],其中 L1 正则化具有一般凸非光滑性质,L2 正则化则是强凸光滑的.损失函数用来控制模型的训练精度,常用的损失函数主要有 Hinge[5]、L2[6]、Logistic[7]、最小二乘[8]等,这几种损失中除 Hinge 外其余均具有光滑性质.在讨论机器学习优化问题的目标函数时,无论是正则化还是损失只要有一个满足强凸性质,则整个目标函数就是强凸的;类似地,如果其中之一满足非光滑性质,则整个目标函数就是非光滑的.光滑性和凸性对优化问题的求解起着举足轻重的作用.

机器学习问题有了统一的研究框架后,多数问题可归结为凸优化问题[1],所以优化方法和优化理论的研究成了机器学习的重要组成部分.常用的优化方法有在线优化[9,10]、随机优化[11]和坐标优化[4~6]等,其中坐标优化能够利用数据的稀疏特性,在处理高维稀疏数据(如文本分类数据)时具有较大的优势[12].在线和随机优化每次迭代仅处理一个样本,计算代价小,特别适合处理机器学习所面临的大规模冗余数据,因此广受学者们青睐.

同一种优化方法若使用的优化理论不同,其结果也会相差甚远.截至目前,机器学习的研究者们已经将多

收稿日期:2014-01-23;修回日期:2014-04-16;责任编辑:梅志强
基金项目:国家自然科学基金(No.61273296,No.60975040);安徽省自然科学基金(No.1308085QF121)

种基于不同优化原理的算法引入进来,并取得了显著的实用效果.例如,比较经典的有随机梯度下降[13](Stochastic Gradient Descent,SGD)、对偶平均[14](Dual Averaging,DA)、镜面下降[15](Mirror Descent,MD)等方法.通过比较,容易发现,这些方法有一个共同点,即都属于黑箱方法.黑箱方法将正则化和损失函数作为一个整体目标进行处理,没有利用机器学习问题的结构信息,特别是处理 L1 正则化问题时,无法获得稀疏解.2009 年,Xiao[16]指出黑箱方法在解决正则化学习问题时缺少发掘问题结构的能力,为此,Xiao 将 Nesterov 的 DA 推广为具有较好稀疏性的 RDA(Regularized Dual Averaging)算法;随后,Duchi 等人[17]对 MD 方法进行改进,提出 COMID 算法,使得学术界很多零散的研究成果得到统一.对于 L1 正则化问题,COMID 和 RDA 在优化过程中将正则化和损失函数分别看待,仅对损失函数进行近似线性展开而保持正则化不变,因此能够得到稀疏解.

RDA 和 COMID 算法不再属于黑箱方法的范畴,更符合机器学习结构优化的框架.虽然两种方法对于一般凸和强凸目标函数均得到 $O(1/\sqrt{T})$ 和 $O(\log T/T)$ 的收敛速率,但通过仔细分析不难发现,RDA 和 COMID 都是先以在线算法的形式被提出,然后才扩展为随机算法.在线算法的理论分析工具为 regret 界,当讨论收敛速率时,往往是指随机算法,RDA 和 COMID 的收敛速率均是由在线算法的 regret 使用 online-to-batch[18]转换得来,此时的收敛速率是指算法所有 T 次迭代平均解的收敛速率,其稀疏性不免差强人意.随机算法单次迭代得到的瞬时解的稀疏性固然好,但目前关于这两种算法的瞬时解的收敛速率缺乏较为完善的理论分析.

在对 RDA 研究的基础上,Chen 等人[19]通过使用加权后的平均梯度代替原来 RDA 的简单平均梯度提出 ORDA(Optimal RDA)算法,在处理非光滑强凸损失函数时该算法不仅能够把 RDA 原来 $O(\log T/T)$ 的收敛速率提升至最优的 $O(1/T)$,而且能够得到瞬时输出的收敛速率,但美中不足的是,ORDA 算法改变了 RDA 的迭代形式,已经不是标准的 RDA 算法.

2013 年,Shamir 等人[20]对 SGD 方法研究时指出,标准 SGD 算法在求解非光滑一般凸和强凸目标函数时,分别能够得到 $O(\log T/\sqrt{T})$ 和 $O(\log T/T)$ 的瞬时收敛速率,并给出了完善的理论证明,该项研究成果首次给出了 SGD 的瞬时收敛速率,同时也在一定程度上解决了文献[21]所提出的 open 问题,意义重大.由于 SGD 是 MD 方法的一种特殊情况[15],并且 COMID 算法是由 MD 演化而来,受文献[20]启发,我们很自然的想到,COMID 方法在处理类似问题时能否得到相同的瞬时收敛速率呢?

本文考虑 L1 正则化非光滑损失的随机优化问题,针对正则化为一般凸和强凸两种情况,在不改变 COMID 算法的前提下对其瞬时收敛速率进行理论分析,指出当正则化项仅为 L1 正则化时能够得到 $O(\log T/\sqrt{T})$ 的瞬时收敛速率,为 L1 + L2 混合正则化时能够得到 $O(\log T/T)$ 的瞬时收敛速率.这两种收敛速率虽然在理论上没有达到平均解的最优收敛速率,即 $O(1/\sqrt{T})$ 和 $O(1/T)$,但在大规模数据库上的实验表明,该收敛速率与最优收敛速率相差无几.此外,从实验数据不难看出,瞬时解的稀疏度明显高于平均解,特别是当处理稀疏性本身就很差的数据库时(如 a9a、covtype),瞬时解的稀疏度是平均解的 3 ~ 4 倍.

2　随机优化和 COMID 算法

为方便理解,对文中所用数学符号作简要说明.假设训练数据库独立同分布,样本表示为:$(\boldsymbol{x}_t, y_t) \in \mathbf{R}^n \times \{-1, +1\}$, $t = 1, \cdots, m$,其中 m 表示样本个数,n 表示样本维数,记 $\xi = (\boldsymbol{x}, y)$ 表示随机抽取的样本.$r(\boldsymbol{w})$ 表示正则化项,$l(\boldsymbol{w}, \xi)$ 表示损失函数,$\boldsymbol{w} \in \Omega$,其中 Ω 为 $\mathbf{R}^n$ 上的闭凸集合,$\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ 表示向量 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 的内积,$\|\cdot\|_1$ 表示 $L1$ 范数 $\|\cdot\|_2$ 和 $\|\cdot\|$ 均表示 $L2$ 范数.

2.1　随机优化

随机优化问题可表示为如下形式:

$$\min_{\boldsymbol{w} \in \Omega} \Phi(\boldsymbol{w}), \text{其中 } \Phi(\boldsymbol{w}) = E_\xi[r(\boldsymbol{w}) + l(\boldsymbol{w}, \xi)]$$

其中 ξ 表示随机抽取的样本,为方便说明记 $\Phi(\boldsymbol{w}, \xi) = r(\boldsymbol{w}) + l(\boldsymbol{w}, \xi)$,随机算法每步迭代仅优化随机抽取的一个样本,其主要的性能评价指标为收敛速率,是指在数学期望下随机算法输出解对应的目标函数值收敛于最优目标函数值的速率[22],若用 $\boldsymbol{w}_o$ 表示算法的输出解,则随机算法收敛速率的数学表达式为 $E[\Phi(\boldsymbol{w}_o)] - \Phi(\boldsymbol{w}^*)$,其中 $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w} \in \Omega} \Phi(\boldsymbol{w})$.

对于随机算法,由于其目标函数是期望形式,因此无法直接优化求解;另一方面由于样本独立同分布,关于单个样本的目标函数 $\Phi(\boldsymbol{w}_t, \xi)$ 的次梯度 $\boldsymbol{g}_t$ 是整个目标函数 $\Phi(\boldsymbol{w}_t)$ 次梯度的无偏估计[11,19~24],即有 $E[\boldsymbol{g}_t] \in \partial\Phi(\boldsymbol{w}_t)$,因此随机算法在形式上表现为每次仅对随机抽取的单个样本进行优化.

当一个随机算法执行 T 次迭代后,标准平均 $\boldsymbol{w}_o = (\boldsymbol{w}_1 + \cdots + \boldsymbol{w}_T)/T$ 是最常用的输出方式,除此之外还有瞬时[20]、suffix 平均[23]、加权平均[24]和多项式衰减平均[20]等输出方式.就 SGD 求解强凸优化问题而言,标准平均和瞬时输出的收敛速率为 $O(\log T/T)$,其余几种平均方式均能够得到 $O(1/T)$ 的收敛速率.

2.2　COMID 算法

COMID 算法是 Duchi 等人对经典 MD 方法的突破性改进.如果样本维数足够大,MD 方法被认为是最优

的一阶方法[17],其单个样本的主要迭代步骤如下:

$$\boldsymbol{w}_{t+1}=\arg\min_{\boldsymbol{w}\in\Omega}\{\eta_t\langle \boldsymbol{g}_t,\boldsymbol{w}-\boldsymbol{w}_t\rangle+B_\varphi(\boldsymbol{w},\boldsymbol{w}_t)\}$$

上式中函数 $B_\varphi(\boldsymbol{w},\boldsymbol{w}_t)$表示 φ 函数的 Bregman Divergence[25],$\boldsymbol{g}_t$ 为目标函数 $r(\boldsymbol{w}_t)+l(\boldsymbol{w}_t,\xi_t)$在 $\boldsymbol{w}_t$ 处的次梯度,η_t 为步长.

由于 MD 把正则化和损失函数作为统一的整体处理,不能充分发挥正则化所起的结构作用,仍属于黑箱方法.而 COMID 方法采取保持正则化不动,仅对损失函数进行近似展开,此时子问题可以解析求解.其主要迭代步骤为:

$$\boldsymbol{w}_{t+1}=\arg\min_{\boldsymbol{w}\in\Omega}\{\eta_t\langle \boldsymbol{g}_t,\boldsymbol{w}\rangle+\eta_t r(\boldsymbol{w})+B_\varphi(\boldsymbol{w},\boldsymbol{w}_t)\}\quad(1)$$

其中 $\boldsymbol{g}_t$ 仅为损失函数 $l(\boldsymbol{w}_t,\xi_t)$的梯度.算法 1 给出了 COMID 算法的主要迭代过程.

算法 1 COMID 算法

1: Input: initialize $\boldsymbol{w}_1=\boldsymbol{0}$.
2: For $t=1$ to T
3: Compute $\boldsymbol{g}_t\in\partial l(\boldsymbol{w}_t,\xi_t)$.
4: Compute $\boldsymbol{w}_{t+1}$ via 公式(1).
5: End for
6: Output: $\boldsymbol{w}_T$ or $\bar{\boldsymbol{w}}_T=(\boldsymbol{w}_1+\cdots+\boldsymbol{w}_T)/T$.

本文主要考虑能够保证算法得到稀疏解的 L1 正则化及一般凸非光滑损失.正则化为一般凸函数时,取 $r(\boldsymbol{w})=\lambda\|\boldsymbol{w}\|_1$时,步长取 $\eta_t=1/\sqrt{t}$.强凸情况下,$r(\boldsymbol{w})=\lambda\|\boldsymbol{w}\|_1+\sigma/2\|\boldsymbol{w}\|^2$,此时 $r(\boldsymbol{w})$满足 σ-强凸,步长取 $\eta_t=1/\sigma t$.参数 λ 控制着输出解的稀疏性,而 σ 取值的大小直接影响模型的分类精度.

3 瞬时收敛速率分析

为方便证明,本文假设 $\boldsymbol{w}_1=0$,$\mathrm{E}[\|\boldsymbol{w}_t\|]\leqslant M$,$\mathrm{E}[\|\boldsymbol{g}_t\|]\leqslant G$,不失一般性,Bregman 函数取 $B_\varphi(\boldsymbol{w},\boldsymbol{w}_t)=1/2\|\boldsymbol{w}-\boldsymbol{w}_t\|^2$,设 $\boldsymbol{w}^*$ 为优化问题的最优解,即 $\boldsymbol{w}^*=\arg\min_{\boldsymbol{w}\in\Omega}\Phi(\boldsymbol{w})$.现给出常用引理 1.

引理 1 设 $\boldsymbol{w}_t$为算法 1 第 t 步迭代的瞬时输出解,若正则化 $r(\boldsymbol{w})$满足 σ-强凸,对任意 $\boldsymbol{w}\in\Omega$,则有如下关系式成立,

$$\Phi(\boldsymbol{w}_t,\xi)-\Phi(w,\xi)\leqslant r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})+\frac{\eta_t}{2}\|\boldsymbol{g}_t\|^2+\frac{1}{2\eta_t}\|\boldsymbol{w}_t-\boldsymbol{w}\|^2-(\frac{1}{2\eta_t}+\frac{\sigma}{2})\|\boldsymbol{w}_{t+1}-\boldsymbol{w}\|^2$$

引理 1 的证明同文献[17]的引理 6,在此不再叙述.当正则化 $r(\boldsymbol{w})$仅为一般凸函数时,即为引理 1 中 $\sigma=0$ 时的特殊情况.

COMID 求解强凸和一般凸优化问题时,其对应平均输出解的收敛速率分别为 $O(\log T/T)$和 $O(1/\sqrt{T})$,这两项结论对证明 COMID 算法的瞬时收敛速率起着至关重要的作用,为此本文给出引理 2 和引理 3.

引理 2 若正则化 $r(\boldsymbol{w})$满足 σ-强凸,且取步长 $\eta_t=1/\sigma t$,则 COMID 算法运行 $T(T>1)$次迭代后,有如下关系成立,

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[\Phi(\boldsymbol{w}_t)]-\Phi(\boldsymbol{w}^*)\leqslant\frac{1}{2T}\Big[\sigma M^2+\frac{G^2}{\sigma}(\log T+1)\Big]$$

证明 由引理 1,令 $\boldsymbol{w}=\boldsymbol{w}^*$,并两边取期望得,

$$\mathrm{E}[\Phi(\boldsymbol{w}_t)-\Phi(\boldsymbol{w}^*)]\leqslant\mathrm{E}[r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})]+\frac{G^2}{2}\eta_t+\frac{1}{2\eta_t}\mathrm{E}[\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2]-(\frac{1}{2\eta_t}+\frac{\sigma}{2})\mathrm{E}[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2]$$

上式对 $t=1,\cdots T$ 求和得,

$$\sum_{t=1}^{T}\mathrm{E}[\Phi(\boldsymbol{w}_t)-\Phi(\boldsymbol{w}^*)]\leqslant\mathrm{E}[r(\boldsymbol{w}_1)-r(\boldsymbol{w}_{T+1})]+\frac{G^2}{2}\sum_{t=1}^{T}\eta_t+\frac{1}{2}\sum_{t=1}^{T}\mathrm{E}[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2](\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}-\sigma)+\frac{\|\boldsymbol{w}_1-\boldsymbol{w}^*\|^2}{2\eta_1}$$

因为 $\mathrm{E}[\|\boldsymbol{w}_t\|]\leqslant M$,所以有 $\|\boldsymbol{w}^*\|^2\leqslant M^2$,取步长 $\eta_t=1/\sigma t$,$\boldsymbol{w}_1=\boldsymbol{0}$,代入整理得,

$$\sum_{t=1}^{T}E[\Phi(\boldsymbol{w}_t)-\Phi(\boldsymbol{w}^*)]\leqslant\frac{\sigma M^2}{2}+\frac{G^2}{2\sigma}(\log T+1)$$

不等式两边同除 T 整理得引理 2.

引理 3 若正则化 $r(\boldsymbol{w})$为一般凸函数,且取步长 $\eta_t=1/\sqrt{t}$,则 COMID 算法运行 $T(T>1)$次迭代后,有如下关系成立,

$$\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}[\Phi(\boldsymbol{w}_t)]-\Phi(\boldsymbol{w}^*)\leqslant\frac{\sqrt{T}}{T}(2M^2+G^2)$$

证明 引理 3 的证明与引理 2 类似,令引理 1 中 $\sigma=0$,$\boldsymbol{w}=\boldsymbol{w}^*$,两边取期望并对 $t=1,\cdots T$ 求和得,

$$\sum_{t=1}^{T}\mathrm{E}[\Phi(\boldsymbol{w}_t)-\Phi(\boldsymbol{w}^*)]\leqslant\mathrm{E}[r(\boldsymbol{w}_1)-r(\boldsymbol{w}_{T+1})]+\frac{G^2}{2}\sum_{t=1}^{T}\eta_t+\frac{1}{2}\sum_{t=2}^{T}\mathrm{E}[\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2](\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}})+\frac{\|\boldsymbol{w}_1-\boldsymbol{w}^*\|^2}{2\eta_1}$$

因为 $\mathrm{E}[\|\boldsymbol{w}_t\|]\leqslant M$,所以 $\mathrm{E}[\|\boldsymbol{w}-\boldsymbol{w}^*\|^2]\leqslant 4M^2$,又 $\eta_t=1/\sqrt{t}$,$\boldsymbol{w}_1=\boldsymbol{0}$,所以有,

$$\sum_{t=1}^{T}\mathrm{E}[\Phi(\boldsymbol{w}_t)-\Phi(\boldsymbol{w}^*)]\leqslant(2M^2+G^2)\sqrt{T}$$

不等式两边同除 T 整理得引理 3.

正则化方法与黑箱方法有着本质的区别,在使用软阈值方法求解时会出现正则化的交错项,为此我们给出引理 4 及引理 5.

引理 4 设向量 $\boldsymbol{w}\in\mathbf{R}^n$,其 $L1$ 范数为:$\|\boldsymbol{w}\|_1=$

$\sum_{i=1}^{n}|w_i|$,$L2$ 范数为:$\|w\|_2=\|w\|=\sqrt{\sum_{i=1}^{n}w_i^2}$,则有如下不等式成立,

$$\|w\|\leqslant\|w\|_1\leqslant\sqrt{n}\|w\|,\ \forall\ w\in\mathbf{R}^n$$

引理 4 给出一个向量的 $L1$、$L2$ 范数之间的关系,其证明比较简单,文献[12]中也有涉及,本文不再证明.在接下来引理 5 的证明中会反复使用引理 4 的结论.

引理 5　当正则化 $r(w)=\lambda\|w\|_1+\sigma/2\|w\|^2$ 时,采用软阈值方法求解子问题式(1),则正则化交错项有如下关系式成立,

$$\mathrm{E}[r(w_t)-r(w_{t+1})]\leqslant A\eta_t^2+B\eta_t$$

其中 $A=\frac{\sigma}{2}\max\{G^2+2\sqrt{n}\lambda G+n\lambda^2,2\sqrt{n}\lambda\sigma M+n\lambda^2+\sigma^2M^2+4\sqrt{n}\lambda G-G^2\}$,$B=2\sqrt{n}\lambda\sigma M+n\lambda^2+\sigma^2M^2+\sqrt{n}\lambda G+\frac{\sigma(M^2+G^2)}{2}$.

证明　当 $r(w)=\lambda\|w\|_1+\sigma/2\|w\|^2$ 时,式(1)所要解决的优化问题为:

$$w_{t+1}=\arg\min_{w\in\Omega}\left\{\frac{\sigma\eta_t+1}{2}\|w\|^2+\langle\eta_t g_t-w_t,w\rangle+\lambda\eta_t\|w\|_1\right\}$$

由于向量 w 的每一维不相关,因此可以将其转换为如下单维优化问题,其中 $1\leqslant j\leqslant n$.

$$w_{t+1,j}=\arg\min_{w_j}\left\{\frac{\sigma\eta_t+1}{2}w_j^2+\langle\eta_t g_{t,j}-w_{t,j},w_j\rangle+\lambda\eta_t|w_j|\right\}$$

上式为“二次项 + 一次项 + L1 正则化项”,使用软阈值方法[16]求得解析解为:

若 $|w_{t,j}-\eta_t g_{t,j}|\leqslant\lambda\eta_t$,则:

$$w_{t+1,j}=0$$

若 $|w_{t,j}-\eta_t g_{t,j}|>\lambda\eta_t$,则:

$$w_{t+1,j}=\frac{1}{\sigma\eta_t+1}[w_{t,j}-\eta_t g_{t,j}-\lambda\eta_t\ \mathrm{sgn}(w_{t,j}-\eta_t g_{t,j})]$$

由此可分为如下 2 种情况讨论.

情况 1　当 $|w_{t,j}-\eta_t g_{t,j}|\leqslant\lambda\eta_t$ 即 $|w_{t,j}|-|\eta_t g_{t,j}|\leqslant\lambda\eta_t$ 时,有 $w_{t+1,j}=0$,此时单维的正则化交错项为:

$$\begin{aligned}r(w_t)_j-r(w_{t+1})_j&=r(w_t)_j=\lambda|w_{t,j}|+\frac{\sigma}{2}w_{t,j}^2\\&\leqslant\lambda\eta_t(|g_{t,j}|+\lambda)+\frac{\sigma\eta_t^2}{2}(|g_{t,j}|+\lambda)^2\\&=\frac{\sigma\eta_t^2}{2}|g_{t,j}|^2+\lambda\eta_t(1+\sigma\eta_t)|g_{t,j}|+n\lambda^2\eta_t(1+\frac{\sigma\eta_t}{2})\end{aligned}$$

所以有,

$$\begin{aligned}r(w_t)-r(w_{t+1})&=\sum_{j=1}^{n}[r(w_t)_j-r(w_{t+1})_j]\\&=\sum_{j=1}^{n}\left[\frac{\sigma\eta_t^2}{2}|g_{t,j}|^2+\lambda\eta_t(1+\sigma\eta_t)|g_{t,j}|+\lambda^2\eta_t(1+\frac{\sigma\eta_t}{2})\right]\\&\leqslant\frac{\sigma\eta_t^2}{2}\|g_t\|_2^2+\lambda\sqrt{n}\eta_t(1+\sigma\eta_t)\|g_t\|_2+n\lambda^2\eta_t(1+\frac{\sigma\eta_t}{2})\end{aligned}$$

对上式两边取期望得,

$$\begin{aligned}&\mathrm{E}[r(w_t)-r(w_{t+1})]\\&\leqslant\frac{\sigma}{2}(G^2+2\sqrt{n}\lambda G+n\lambda^2)\eta_t^2+(\sqrt{n}\lambda G+n\lambda^2)\eta_t\end{aligned}$$

令 $A_1=\frac{\sigma}{2}(G^2+2\sqrt{n}\lambda G+n\lambda^2)$,$B_1=\sqrt{n}\lambda G+n\lambda^2$.

则有:$\mathrm{E}[r(w_t)-r(w_{t+1})]\leqslant A_1\eta_t^2+B_1\eta_t$.

情况 2　当 $|w_{t,j}-\eta_t g_{t,j}|>\lambda\eta_t$ 时,有 $w_{t+1,j}=\frac{1}{\sigma\eta_t+1}[w_{t,j}-\eta_t g_{t,j}-\lambda\eta_t\ \mathrm{sgn}(w_{t,j}-\eta_t g_{t,j})]$.为方便,使用 sgn 作为 $\mathrm{sgn}(w_{t,j}-\eta_t g_{t,j})$的简写,此时有,

$$\begin{aligned}&r(w_t)_j-r(w_{t+1})_j=\lambda|w_{t,j}|-\lambda|w_{t+1,j}|+\frac{\sigma}{2}w_{t,j}^2-\frac{\sigma}{2}w_{t+1,j}^2\\&\leqslant\lambda|w_{t,j}-\frac{1}{\sigma\eta_t+1}(w_{t,j}-\eta_t g_{t,j}-\lambda\eta_t\ \mathrm{sgn})|\\&+\frac{\sigma}{2}\left[|w_{t,j}|^2-\frac{1}{(\sigma\eta_t+1)^2}|w_{t,j}-\eta_t g_{t,j}-\lambda\eta_t\ \mathrm{sgn}|^2\right]\\&\leqslant\frac{\lambda\eta_t}{\sigma\eta_t+1}(\sigma|w_{t,j}|+|g_{t,j}|+\lambda)+\frac{\sigma}{2(\sigma\eta_t+1)^2}\\&\cdot[(\sigma\eta_t+1)^2|w_{t,j}|^2-(|w_{t,j}|-\eta_t|g_{t,j}+\lambda\mathrm{sgn}|)^2]\\&\leqslant\frac{\lambda\eta_t}{\sigma\eta_t+1}(\sigma|w_{t,j}|+|g_{t,j}|+\lambda)+\frac{\sigma\eta_t}{2(\sigma\eta_t+1)^2}[\sigma(\sigma\eta_t+2)\\&\cdot|w_{t,j}|^2+2|w_{t,j}||g_{t,j}|+2\lambda|w_{t,j}|-\eta_t(|g_{t,j}|-\lambda)^2]\\&=\frac{\lambda\eta_t}{\sigma\eta_t+1}(\sigma|w_{t,j}|+|g_{t,j}|+\lambda)+\frac{\sigma\eta_t}{2(\sigma\eta_t+1)^2}[(\sigma^2\eta_t+2\sigma+1)\\&\cdot|w_{t,j}|^2+|g_{t,j}|^2+2\lambda|w_{t,j}|-\eta_t(|g_{t,j}|^2-2\lambda|g_{t,j}|+\lambda^2)]\\&=\frac{\lambda\eta_t}{\sigma\eta_t+1}(\sigma|w_{t,j}|+|g_{t,j}|+\lambda)+\frac{\sigma\eta_t}{2(\sigma\eta_t+1)^2}[(\sigma^2\eta_t+2\sigma+1)\\&\cdot|w_{t,j}|^2+2\lambda|w_{t,j}|+(1-\eta_t)|g_{t,j}|^2+2\lambda\eta_t|g_{t,j}|-\lambda^2\eta_t]\end{aligned}$$

所以有,

$$\begin{aligned}r(w_t)-r(w_{t+1})&=\sum_{j=1}^{n}[r(w_t)_j-r(w_{t+1})_j]\\&=\frac{\lambda\eta_t}{\sigma\eta_t+1}(\sigma\|w_t\|_1+\|g_t\|_1+n\lambda)+\frac{\sigma\eta_t}{2(\sigma\eta_t+1)^2}\\&[(\sigma^2\eta_t+2\sigma+1)\|w_t\|^2+2\lambda\|w_t\|_1+(1-\eta_t)\\&\cdot\|g_t\|^2+2\lambda\eta_t\|g_t\|_1-n\lambda^2\eta_t]\end{aligned}$$

两边取期望并整理得,

$$\begin{aligned}&\mathrm{E}[r(w_t)-r(w_{t+1})]\leqslant\frac{\lambda\eta_t}{\sigma\eta_t+1}(\sqrt{n}\sigma M+\sqrt{n}G+n\lambda)+\\&\frac{\sigma\eta_t}{2(\sigma\eta_t+1)^2}[(\sigma^2\eta_t+2\sigma+1)M^2+2\sqrt{n}\lambda M+(1-\eta_t)G^2+\\&2\eta_t\sqrt{n}\lambda G-n\lambda^2\eta_t]\\&\leqslant\frac{\sigma}{2}(\sigma^2M^2+2\sqrt{n}\lambda\sigma M+4\sqrt{n}\lambda G+n\lambda^2-G^2)\eta_t^2\\&+\left[\sigma^2M^2+2\sqrt{n}\lambda\sigma M+\sqrt{n}\lambda G+n\lambda^2+\frac{\sigma(M^2+G^2)}{2}\right]\eta_t\end{aligned}$$

令 $A_2=\frac{\sigma}{2}(\sigma^2M^2+2\sqrt{n}\lambda\sigma M+4\sqrt{n}\lambda G+n\lambda^2-G^2)$

$$B_2=\sigma^2M^2+2\sqrt{n}\lambda\sigma M+\sqrt{n}\lambda G+n\lambda^2+\frac{\sigma(M^2+G^2)}{2}$$

则有 $\mathrm{E}[r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})]\leqslant A_2\eta_t^2+B_2\eta_t$.

综合以上 2 种情况,令 $A=\max\{A_1,A_2\}$, $B=B_2$,则有,$\mathrm{E}[r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})]\leqslant A\eta_t^2+B\eta_t$

故引理 5 证毕.

下面给出通过引理 5 得到的一个推论.

推论 当正则化 $r(\boldsymbol{w})=\lambda\|\boldsymbol{w}\|_1$时,即引理 5 中 $\sigma=0$ 的特殊情况,此时正则化交错项满足关系式 $\mathrm{E}[r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})]\leqslant(n\lambda^2+\sqrt{n}\lambda G)\eta_t$.

根据引理 1、引理 2 及引理 5,我们首先给出正则化项 $r(\boldsymbol{w})$满足 σ-强凸时,COMID 的瞬时收敛速率.

定理 1 若 $r(\boldsymbol{w})=\lambda\|\boldsymbol{w}\|_1+\sigma/2\|\boldsymbol{w}\|^2$,且取步长 $\eta_t=1/\sigma t$,则 COMID 算法运行 $T(T>1)$次迭代后,其瞬时收敛速率满足如下关系,

$$\mathrm{E}[\Phi(\boldsymbol{w}_T)]-\Phi(\boldsymbol{w}^*)\leqslant\frac{1}{2\sigma^2T}[(4A+4\sigma B+3\sigma G^2)\log T+2(A+\sigma B+\sigma G^2)+\sigma^3M^2]$$

其中参数 A 和 B 同引理 5.

证明 根据引理 1,并对两边同时取期望得,

$$\mathrm{E}[\Phi(\boldsymbol{w}_t)-\Phi(\boldsymbol{w})]\leqslant\mathrm{E}[r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})]+\frac{G^2}{2}\eta_t+\frac{1}{2\eta_t}\mathrm{E}[\|\boldsymbol{w}_t-\boldsymbol{w}\|^2]-(\frac{1}{2\eta_t}+\frac{\sigma}{2})\mathrm{E}[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}\|^2]$$

令 $k\in\{1,2,\cdots,T-1\}$,并对 $t=T-k,\cdots,T$ 求和得,

$$\sum_{t=T-k}^{T}\mathrm{E}[\Phi(\boldsymbol{w}_t)-\Phi(\boldsymbol{w})]\leqslant\sum_{t=T-k}^{T}\mathrm{E}[r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})]+\frac{G^2}{2}\sum_{t=T-k}^{T}\eta_t+\frac{1}{2}\sum_{t=T-k}^{T}\mathrm{E}[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}\|^2](\frac{1}{\eta_{t+1}}-\frac{1}{\eta_t}-\sigma)+\frac{\|\boldsymbol{w}_{T-k}-\boldsymbol{w}\|^2}{2\eta_{T-k}}-\frac{\|\boldsymbol{w}_{T+1}-\boldsymbol{w}\|^2}{2\eta_{T+1}}$$

根据引理 5 知:$\mathrm{E}[r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})]\leqslant A\eta_t^2+B\eta_t$,并令 $\eta_t=1/\sigma t$, $\boldsymbol{w}=\boldsymbol{w}_{T-k}$,代入并化简得,

$$\sum_{t=T-k}^{T}-\mathrm{E}[\Phi(\boldsymbol{w}_t)]-(k+1)\mathrm{E}[\Phi(\boldsymbol{w}_{T-k})]\leqslant\frac{A}{\sigma^2}\sum_{t=T-k}^{T}\frac{1}{t^2}+\frac{2B+G^2}{2\sigma}\sum_{t=T-k}^{T}\frac{1}{t}$$

两边同除以 $k+1$,并令 $S_k=\frac{1}{k+1}\sum_{t=T-k}^{T}\mathrm{E}[\Phi(\boldsymbol{w}_t)]$,所以有,

$$\mathrm{E}[S_k)]-\mathrm{E}[\Phi(\boldsymbol{w}_{T-k})]\leqslant\frac{A}{\sigma^2}\sum_{t=T-k}^{T}\frac{1}{(k+1)t^2}+\frac{2B+G^2}{2\sigma}\sum_{t=T-k}^{T}\frac{1}{(k+1)t}$$

又因为 $\mathrm{E}[\Phi(\boldsymbol{w}_{T-k})]=(k+1)\mathrm{E}[S_k]-k\mathrm{E}[S_{k-1}]$,所以,

$$\mathrm{E}[S_{k-1}]\leqslant\mathrm{E}[S_k]+\frac{A}{\sigma^2}\sum_{t=T-k}^{T}\frac{1}{k(k+1)t^2}+\frac{2B+G^2}{2\sigma}\sum_{t=T-k}^{T}\frac{1}{k(k+1)t}$$

重复使用该递推式,并对 $k=1$ 到 $T-1$ 求和得,

$$\mathrm{E}[S_0]\leqslant\mathrm{E}[S_{T-1}]+\frac{A}{\sigma^2}\sum_{k=1}^{T-1}\sum_{t=T-k}^{T}\frac{1}{k(k+1)t^2}+\frac{2B+G^2}{2\sigma}\sum_{k=1}^{T-1}\sum_{t=T-k}^{T}\frac{1}{k(k+1)t}$$

因为 $\sum_{t=T-k}^{T}\frac{1}{t^2}\leqslant\frac{1}{(T-k)^2}+\int_{T-k}^{T}\frac{1}{t^2}\,\mathrm{d}t\leqslant\frac{1}{T-k}+\frac{k}{T(T-k)}$,并且 $\sum_{t=T-k}^{T}\frac{1}{t}\leqslant\frac{k+1}{T-k}$,所以有,

$$\sum_{k=1}^{T-1}\sum_{t=T-k}^{T}\frac{1}{k(k+1)t}\leqslant\sum_{k=1}^{T-1}\frac{1}{k(T-k)}\leqslant\frac{2\log T+1}{T}$$

$$\sum_{k=1}^{T-1}\sum_{t=T-k}^{T}\frac{1}{k(k+1)t^2}\leqslant\sum_{k=1}^{T-1}\Big[\frac{1}{k(k+1)}\cdot\frac{1}{T-k}+\frac{1}{T(k+1)(T-k)}\Big]\leqslant\sum_{k=1}^{T-1}\frac{1}{k(T-k)}\leqslant\frac{2\log T+1}{T}$$

根据 S_k定义知,$\mathrm{E}[S_0]=\mathrm{E}[\Phi(\boldsymbol{w}_T)]$,由引理 2 知,$\mathrm{E}[S_{T-1}]\leqslant\Phi(\boldsymbol{w}^*)+\frac{1}{2T}\Big[\sigma M^2+\frac{G^2}{\sigma}(\log T+1)\Big]$,代入并整理得,

$$\mathrm{E}[\Phi(\boldsymbol{w}_T)]-\Phi(\boldsymbol{w}^*)\leqslant\frac{1}{2\sigma^2T}[(4A+4\sigma B+3\sigma G^2)\log T+2(A+\sigma B+\sigma G^2)+\sigma^3M^2]$$

定理 1 证毕.

类似地,由引理 1、引理 3 及推论很容易得到正则化 $r(\boldsymbol{w})$为一般凸函数时 COMID 算法的瞬时收敛速率,由此我们给出定理 2.

定理 2 若 $r(\boldsymbol{w})=\lambda\|\boldsymbol{w}\|_1$,且取步长 $\eta_t=1/\sqrt{t}$,则 COMID 算法运行 $T(T>1)$次迭代后,其瞬时收敛速率满足如下关系,

$$\mathrm{E}[\Phi(\boldsymbol{w}_T)]-\Phi(\boldsymbol{w}^*)\leqslant\frac{1}{\sqrt{T}}[2M^2+G^2+(2n\lambda^2+2\sqrt{n}\lambda G+G^2+2M^2)(1+\log T)]$$

证明 与定理 1 证明类似,令引理 1 中 $\sigma=0$,两边同时取期望并对 $t=T-k,\cdots T$ 求和得,

$$\mathrm{E}[\Phi(\boldsymbol{w}_t)-\Phi(\boldsymbol{w})]\leqslant\sum_{t=T-k}^{T}\mathrm{E}[r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})]+\frac{1}{2}\sum_{t=T-k}^{T}\mathrm{E}[\|\boldsymbol{w}_t-\boldsymbol{w}\|^2](\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}})+\frac{\|\boldsymbol{w}_{T-k}-\boldsymbol{w}\|^2}{2\eta_{T-k-1}}-\frac{\|\boldsymbol{w}_{T+1}-\boldsymbol{w}\|^2}{2\eta_T}+\frac{G^2}{2}\sum_{t=T-k}^{T}\eta_t$$

由 $\mathrm{E}[\|\boldsymbol{w}_t\|]\leqslant M$,易知 $\mathrm{E}[\|\boldsymbol{w}_t-\boldsymbol{w}_{T-k}\|^2]\leqslant4M^2$,由推论知 $\mathrm{E}[r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})]\leqslant(n\lambda^2+\sqrt{n}\lambda G)\eta_t$,并令 $\eta_t=1/\sqrt{t}$, $\boldsymbol{w}=\boldsymbol{w}_{T-k}$代入得,

$$\sum_{t=T-k}^{T}\mathrm{E}[\Phi(\boldsymbol{w}_t)-\Phi(\boldsymbol{w}_{T-k})]$$

$$\leqslant 2(n\lambda^2+\sqrt{n}\lambda G+G^2/2)(\sqrt{T}-\sqrt{T-k})$$
$$+2M^2(\sqrt{T}-\sqrt{T-k-1})$$
$$\leqslant(2n\lambda^2+2\sqrt{n}\lambda G+G^2+2M^2)(\sqrt{T}-\sqrt{T-k-1})$$
$$=(2n\lambda^2+2\sqrt{n}\lambda G+G^2+2M^2)\frac{k+1}{\sqrt{T}+\sqrt{T-k-1}}$$
$$\leqslant(2n\lambda^2+2\sqrt{n}\lambda G+G^2+2M^2)\frac{k+1}{\sqrt{T}}$$

同理,令 $S_k=\frac{1}{k+1}\sum_{t=T-k}^{T}\mathrm{E}[\Phi(\boldsymbol{w}_t)]$,两边同除以 $k+1$,

$$\mathrm{E}[S_k]-\Phi(\boldsymbol{w}_{T-k})\leqslant(2n\lambda^2+2\sqrt{n}\lambda G+G^2+2M^2)\frac{1}{\sqrt{T}}$$

因为 $\mathrm{E}[\Phi(\boldsymbol{w}_{T-k})]=(k+1)\mathrm{E}[S_k]-k\mathrm{E}[S_{k-1}]$,所以,

$$\mathrm{E}[S_{k-1}]\leqslant\mathrm{E}[S_k]+(2n\lambda^2+2\sqrt{n}\lambda G+G^2+2M^2)\frac{1}{\sqrt{T}}\cdot\frac{1}{k}$$

重复使用上式并从 $k=1$ 到 $T-1$ 求和得,

$$\mathrm{E}[S_0]\leqslant\mathrm{E}[S_{T-1}]+(2n\lambda^2+2\sqrt{n}\lambda G+G^2+2M^2)\frac{1}{\sqrt{T}}\cdot\sum_{k=1}^{T-1}\frac{1}{k}$$

由 S_k 定义知 $\mathrm{E}[S_0]=\mathrm{E}[\Phi(\boldsymbol{w}_T)]$,且 $\sum_{k=1}^{T-1}\frac{1}{k}\leqslant 1+\log T$,由引理 3 知 $\mathrm{E}[S_{T-1}]\leqslant\Phi(\boldsymbol{w}^*)+(2M^2+G^2)\frac{\sqrt{T}}{T}$.

所以,$\mathrm{E}[\Phi(\boldsymbol{w}_T)]-\Phi(\boldsymbol{w}^*)\leqslant\frac{1}{\sqrt{T}}[2M^2+G^2+(2n\lambda^2+2\sqrt{n}\lambda G+G^2+2M^2)(1+\log T)]$

定理 2 证毕.

从定理 1 和定理 2 不难看出,COMID 算法在求解强凸和一般凸非光滑随机优化问题时,分别能够得到 $O(\log T/T)$和 $O(\log T/\sqrt{T})$的瞬时收敛速率.

4　数值实验

本节的主要目的是在大规模数据库上通过实验对 COMID 算法瞬时输出解的实际效果进行验证.实验环境为 Oracle Sun Fire X4170 M2 服务器,配置为,Oracle Solaris 操作系统,2.40GHz Intel(R) Xeon(R) CPU,12GB 内存,实验平台为 LIBLINEAR[26].

4.1　实验数据库及算法描述

本文实验所采用的 4 个大规模数据库分别为 a9a、CCAT、astro-physic 和 covtype.表 1 为 3 个数据库的详细描述.

表 1　实验数据库描述

数据库	训练样本数	测试样本数	维数
a9a	24,703	7,858	123
CCAT	23,149	781,265	47,236
astro-physic	29,882	32,487	99,757
covtype	522,911	58,101	54

实验比较 3 种不同类型的算法,即 COMID、SGD 和 RDA.实验中算法后缀"ii"为 individual iterates 的首字母缩写,表示算法取瞬时输出;同理,"ave"表示平均输出(average),计算方式为 $\boldsymbol{w}=(\boldsymbol{w}_1+\boldsymbol{w}_2+\cdots+\boldsymbol{w}_T)/T$;"wei"表示加权平均输出[24].算法前的 L1 和 L2 表示算法所使用的正则化,其中 L1L2COMID-ii 表示使用 L1 + L2 混合正则化.试验中所有算法的损失函数均采用非光滑 Hinge 损失,即 $l(\boldsymbol{w},\xi)=\max\{0,1-y\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}\}$.

4.2　实验方法及结论

实验过程中,我们对每个数据库中样本采取随机抽取的方式,算法进行 10000 步迭代后终止.为公平起见,算法中的参数均在 $10^{-6}\sim10^{2}$范围内采用网格搜索方式取最优参数,并且算法在每个数据库上运行 10 次,每一步迭代的结果均取 10 次结果的平均值和方差.其中,表 2 和表 3 为算法第 10000 步迭代后的测试错误率和稀疏度的平均值和方差,图 1 和图 2 表示算法迭代过程中的稀疏度变化图和目标函数收敛速率图,图中为每 100 步迭代取一个记录点,每个记录点的结果为 10 次结果的平均值.限于篇幅,对于比较图我们仅给出 a9a、CCAT 和 astro-physic 这 3 个数据库的实验结果.

表 2 给出每个算法在 4 个数据库上的测试错误率和方差.测试错误率越小,表示算法在测试库上取得的正确率越高.从表中可以看出,L1L2COMID-ii 算法比其他几种算法的正确率要低一些,但总的看来,这几种算法所取得的正确率基本相同,没有明显差距.此外,从各算法的方差容易看出,算法取瞬时解没有取平均解的稳定性好,这一现象在 a9a 和 covtype 数据库上表现较为明显.

表 2　测试错误率和方差

算法	a9a	CCAT	astro-physic	covtype
L1L2COMID-ii	0.1671 ± 0.0096	0.0918 ± 0.0012	0.0452 ± 0.0017	0.2436 ± 0.0065
L1COMID-ii	0.1535 ± 0.0019	0.0890 ± 0.0009	0.0465 ± 0.0010	0.2345 ± 0.0031
L1COMID-ave	0.1534 ± 0.0012	0.0992 ± 0.0013	0.0516 ± 0.0019	0.2354 ± 0.0023
L2SGD-ii	0.1569 ± 0.0034	0.0866 ± 0.0020	0.0402 ± 0.0008	0.2391 ± 0.0044
L2SGD-wei	0.1534 ± 0.0013	0.0865 ± 0.0010	0.0405 ± 0.0006	0.2357 ± 0.0014
L1RDA-ave	0.1513 ± 0.0010	0.0786 ± 0.0007	0.0407 ± 0.0005	0.2329 ± 0.0017

表 3 稀疏度和方差

算法	a9a	CCAT	astro-physic	covtype
L1L2COMID-ii	0.2553 ± 0.0290	0.7126 ± 0.0052	0.8737 ± 0.0020	0.4389 ± 0.0290
L1COMID-ii	0.2358 ± 0.0287	0.8364 ± 0.0046	0.9324 ± 0.0019	0.4537 ± 0.0306
L1COMID-ave	0.0699 ± 0.0172	0.6422 ± 0.0073	0.8578 ± 0.0026	0.1019 ± 0.0200
L2SGD-ii	0.0659 ± 0.0146	0.4724 ± 0.0032	0.8368 ± 0.0025	0.1037 ± 0.0199
L2SGD-wei	0.0626 ± 0.0144	0.4713 ± 0.0030	0.8340 ± 0.0030	0.0778 ± 0.0210
L1RDA-ave	0.0829 ± 0.0236	0.6319 ± 0.0036	0.8269 ± 0.0033	0.3630 ± 0.0329

表 3 为算法在各数据库上的稀疏度和方差. 稀疏度指算法输出解向量中零维所占的比例,稀疏度越高表示算法稀疏性越好. 从比较结果可以看出,算法 L1L2COMID-ii 和 L1COMID-ii 的稀疏性明显好于其他几种算法,L2SGD-ii 算法虽然也取瞬时解,但由于其使用 L2 正则化且是黑箱方法,故稀疏性较差. 从 L1COMID-ii 和 L1COMID-ave 算法在 a9a 和 covtype 数据库上的实验数据可以看出,瞬时解的稀疏度甚至是平均解的 3 ~ 4 倍. 比较 L1COMID-ii 和 L1L2COMID-ii 算法可以发现,前者的稀疏性较好,主要是因为后者使用的 L1 + L2 混合正则化使解的非零维增加,从而影响了解的稀疏性.

为更好的说明算法在迭代过程中的稀疏性变化,我们给出各算法的稀疏度比较图,如图 1 所示. 横纵坐标分别为迭代次数和稀疏度,从图中不难发现,在经过较少的迭代次数后,L1L2COMID-ii 和 L1COMID-ii 算法的稀疏性一直优于其他算法.

图 2 为比较算法的收敛速率比较图. 图中横纵坐标轴均为对数坐标轴,横坐标表示迭代次数,纵坐标表示

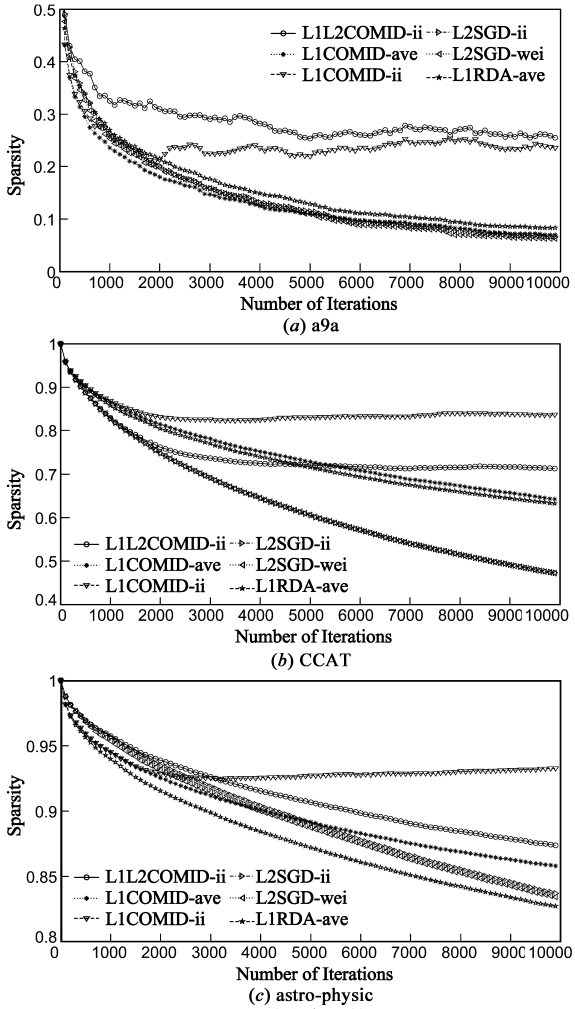


图1 稀疏度比较图

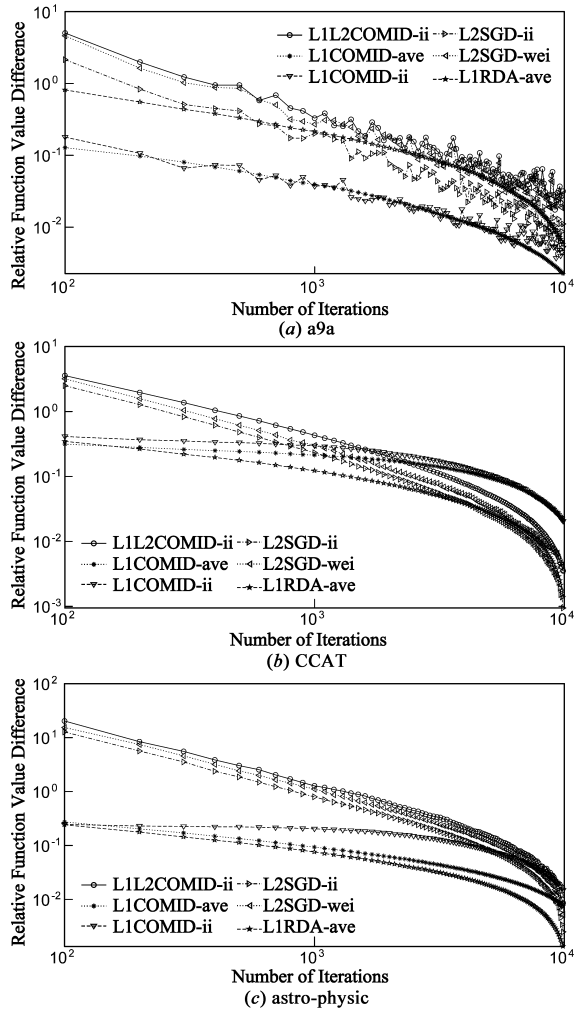


图2 目标函数收敛速率比较图

相对目标函数(Relative Function Value Difference),即当前目标函数值与最优目标函数值之差,其数学表达式为 $\Phi(\boldsymbol{w})-\Phi(\boldsymbol{w}^*)$,具体计算时最优目标函数值取所有迭代中最小的目标函数值.

对于对数坐标轴来说,目标函数下降曲线的斜率对应着收敛速率的阶.我们已经知道算法 L1COMID-ave 和 L1RDA-ave 的收敛速率为 $O(1/\sqrt{T})$,而 L2SGD-ii 算法为 $O(\log T/T)$,L2SGD-wei 算法为 $O(1/T)$,从图 2 可以看出,L1L2COMID-ii、L2SGD-ii 和 L2SGD-wei 这 3 种算法的目标函数下降曲线的斜率基本一致,说明 L1L2COMID-ii 虽然仅得到 $O(\log T/T)$的收敛速率,但考虑系数的情况下,其收敛速率非常接近 $O(1/T)$.同理,L1COMID-ii 算法虽具有 $O(\log T/\sqrt{T})$的收敛速率,但与 $O(1/\sqrt{T})$相差不大.

5　总结与展望

本文针对 COMID 算法求解 L1 正则化非光滑损失的随机优化问题,指出当正则化为一般凸时能够得到 $O(\log T/\sqrt{T})$的瞬时收敛速率,当正则化满足强凸性质时能够得到 $O(\log T/T)$的瞬时收敛速率,最后通过实验对算法的性能作了验证.

对于一般凸和强凸情况下 COMID 算法的瞬时收敛速率与平均解的最优收敛速率还存在一定差距,能否在理论上把瞬时收敛速率提升至最优是我们下步主要研究的问题.

参考文献

[1] 陶卿,朱烨雷,等.一种基于 Comid 的非光滑损失随机坐标下降方法[J].电子学报,2013,41(4):768 – 775.
Tao Qing, Zhu Ye-lei, et al. A new comid-based stochastic coordinate descent method for non-smooth losses[J]. Acta Electronica Sinica, 2013, 41(4): 768 – 775. (in Chinese)

[2] Shalev-Shwartz S, Tewari A. Stochastic methods for l1-regularized loss minimization[J]. Journal of Machine Learning Research, 2011, 12(Jun): 1865-1892.

[3] 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.

[4] Shalev-Shwartz S, Zhang T. Proximal stochastic dual coordinate ascent[OL]. http://arxiv.org/abs/1211.2717, 2012-11-12/2012-12-01.

[5] Hsieh C J, Chang K W, et al. A dual coordinate descent method for large-scale linear SVM[A]. Proceedings of the 25th International Conference on Machine Learning[C]. USA: ACM Press, 2008. 408 – 415.

[6] Chang K W, Hsieh C J, Lin C J. Coordinate descent method for large-scale l2-loss linear support vector machines[J]. Journal of Machine Learning Research, 2008, 9(Jun): 1369-1398.

[7] Lin C J, Weng R C, Keerthi S S. Trust region newton method for logistic regression[J]. Journal of Machine Learning Research, 2008, 9(Apr): 627 – 650.

[8] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society, 1996, 58(1): 267 – 288.

[9] Zinkevich M. Online convex programming and generalized infinitesimal gradient ascent[A]. Proceedings of the 20th International Conference on Machine Learning[C]. USA: ACM Press, 2003. 928 – 936.

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

[11] Nemirovski A, Juditsky A, Lan G, et al. Robust stochastic approximation approach to stochastic programming[J]. SIAM Journal on Optimization, 2009, 19(4): 1574 – 1609.

[12] Yuan G X, Chang K W, Hsieh C J, et al. A comparison of optimization methods and software for large-scale l1-regularized linear classification[J]. Journal of Machine Learning Research, 2010, 11(Nov): 3183-3234.

[13] H Robbins, S Monro. A stochastic approximation method[J]. The Annuals of Mathematical Statistics, 1951, 22(3): 400 – 407.

[14] 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.

[15] Beck A, Teboulle M. Mirror descent and nonlinear projected subgradient methods for convex optimization[J]. Operations Research Letters, 2003, 31(3): 167 – 175.

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

[17] Duchi J, Shalev-Shwartz S, Singer Y, Tewari A. Composite objective mirror descent[A]. Proceedings of the 23rd Annual Workshop on Computational Learning Theory[C]. USA: ACM Press, 2010. 116-128.

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

[19] Chen X, Lin Q, Pena J. Optimal regularized dual averaging methods for stochastic optimization[A]. Advances in Neural Information Processing Systems[C]. USA: ACM Press, 2012. 404 – 412.

[20] Shamir O, Zhang T. Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes[A]. Proceedings of the 30th International Conference on Machine Learning[C]. USA: ACM Press, 2013. 71 – 79.

[21] Shamir O. Is averaging needed for strongly convex stochastic gradient descent[A]. Proceedings of the 25th Annual Conference on Learning Theory. JMLR W & CP 23[C]. Edinburgh. Scotland: JMLR, 2012. 1 – 47.

[22] Hazan E, Kale S. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization [A]. Proceedings of the 24th Annual Conference on Learning Theory, JMLR W & CP 19[C]. Budapest, Hungary: JMLR, 2011. 421 – 436.

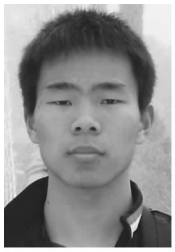
[23] Rakhlin A, Shamir O, Sridharan K. Makinggradient descent optimal for strongly convex stochastic optimization[A]. Proceedings of the 29th International Conference on Machine Learning[C]. USA: ACM Press, 2012. 449 – 456.

[24] Lacoste-Julien S, Schmidt M, Bach F. A simpler approach to obtaining an o(1/t) convergence rate for projected stochastic subgradient descent[OL]. http://arxiv. org/abs/1212. 2002, 2012-12-20/2013-03-01.

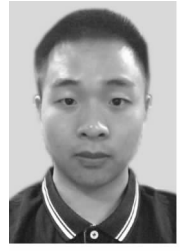
[25] Bregman L M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming[J]. USSR Computational Mathematics and Mathematical Physics, 1967, 7(3): 200 – 217.

[26] Fan R E, Chang K W, Hsieh C J, et al. LIBLINEAR: A library for large linear classification[J]. Journal of Machine Learning Research, 2008, 9: 1871 – 1874.

作者简介



姜纪远 男, 1989 年生于安徽涡阳, 硕士研究生, 主要研究方向为机器学习、凸优化及模式识别.
E-mail: jyjianggle@gmail.com



邵言剑 男, 1990 年生于江苏镇江, 硕士研究生, 主要研究方向为模式识别与人工智能.
E-mail: shy.jian@gmail.com

陶 卿 男, 1965 生于安徽合肥, 博士, 教授, CCF 高级会员, 主要研究领域为机器学习、模式识别及应用数学.
E-mail: taoqing@gmail.com

汪群山 男, 1976 年生于安徽合肥, 硕士, 主要研究方向为机器学习与图像处理.
E-mail: pbxymtn@163.com