软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(9):2160−2171 [doi: 10.13328/j.cnki.jos.004633]

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

一种求解强凸优化问题的最优随机算法*

邵言剑, 陶 卿, 姜纪远, 周 柏

(中国人民解放军陆军军官学院 十一系,安徽 合肥 230031)

通讯作者: 邵言剑, E-mail: shy.jian@gmail.com

摘 要: 随机梯度下降(SGD)算法是处理大规模数据的有效方法之一.黑箱方法 SGD 在强凸条件下能达到最优的 $O(1/T)$收敛速率,但对于求解 $L1+L2$ 正则化学习问题的结构优化算法,如 COMID(composite objective mirror descent)仅具有 $O(\ln T/T)$的收敛速率.提出一种能够保证稀疏性基于 COMID 的加权算法,证明了其不仅具有 $O(1/T)$的收敛速率,还具有 on-the-fly 计算的优点,从而减少了计算代价.实验结果表明了理论分析的正确性和所提算法的有效性.

关键词: 机器学习;随机优化;强凸问题;混合正则化项;COMID (composite objective mirror descent)

中图法分类号: TP301

中文引用格式: 邵言剑,陶卿,姜纪远,周柏.一种求解强凸优化问题的最优随机算法.软件学报,2014,25(9):2160−2171. http://www.jos.org.cn/1000-9825/4633.htm

英文引用格式: Shao YJ, Tao Q, Jiang JY, Zhou B. Stochastic algorithm with optimal convergence rate for strongly convex optimization problems. Ruan Jian Xue Bao/Journal of Software, 2014,25(9):2160−2171 (in Chinese). http://www.jos.org.cn/1000-9825/4633.htm

Stochastic Algorithm with Optimal Convergence Rate for Strongly Convex Optimization Problems

SHAO Yan-Jian, TAO Qing, JIANG Ji-Yuan, ZHOU Bai

(11st Department, Army Officer Academy of PLA, Hefei 230031, China)

Corresponding author: SHAO Yan-Jian, E-mail: shy.jian@gmail.com

Abstract: Stochastic gradient descent (SGD) is one of the efficient methods for dealing with large-scale data. Recent research shows that the black-box SGD method can reach an $O(1/T)$ convergence rate for strongly-convex problems. However, for solving the regularized problem with $L1$ plus $L2$ terms, the convergence rate of the structural optimization method such as COMID (composite objective mirror descent) can only attain $O(\ln T/T)$. In this paper, a weighted algorithm based on COMID is presented, to keep the sparsity imposed by the $L1$ regularization term. A prove is provided to show that it achieves an $O(1/T)$ convergence rate. Furthermore, the proposed scheme takes the advantage of computation on-the-fly so that the computational costs are reduced. The experimental results demonstrate the correctness of theoretic analysis and effectiveness of the proposed algorithm.

Key words: machine learning; stochastic optimization; strongly-convex; hybrid regularization; COMID (composite objective mirror descent)

机器学习面临的数据规模正变得越来越大,比如,一个普通文本数据库就会达到 10^7 样本个数或 10^9 样本维数的规模[1].在对大数据进行处理和分析中可以挖掘出有价值的信息,进而有效地解决或缓解各领域面临的问题.然而,大数据在带来机遇的同时,也面临着众多的困难和挑战,寻求一种高速、有效的计算技术,是目前亟需解决的科学问题.机器学习方法在该背景下得到广泛的应用,其核心问题之一是求解优化问题[2],与传统批处理算法不同的是,随机梯度下降(SGD)算法[3,4]在每次迭代计算时,仅仅优化单个样本点造成的损失,极大地减少了内

* 基金项目: 国家自然科学基金(61273296)
收稿时间: 2014-01-23; 定稿时间: 2014-04-09

存开销,但也正是每次仅仅优化单个样本点造成的损失,收敛速率不可避免地会受到影响.但机器学习问题有其特殊性,即样本集独立同分布,并且通常冗余度较大,仅需要很少一部分样本就能达到所需的泛化能力,因此,迭代部分样本点步骤后,优化问题解的学习精度就已经趋于稳定.特别是在 2007 年,Shalev-Shwarts 等人[5]在投影次梯度方法的基础上得到了求解支持向量机问题的 SGD,该方法又称为 Pegasos.Pegasos 在处理 80 万个样本的路透社文本 RCV1 数据库仅需几秒的时间,与当时主流的大规模算法相比,SGD 在处理实际学习问题中具有明显的优势.因此,SGD 算法无论是在理论还是在实用上成为求解大规模学习问题的有效方法.

对于强凸优化问题,Pegasos 的收敛速率只达到 $O(\log T/T)$,与理论上的最优收敛速率 $O(1/T)$存在着明显的差距.为解决这一问题,Hazan 等人[6]在 SGD 方法中嵌入内循环,提出 EPOCH-GD 随机算法,获得了 $O(1/T)$的最优收敛速率.但 Rakhlin 等人[7]认为,EPOCH-GD 算法与标准的 SGD 在算法形式上存在较大差别,因此不能说明 SGD 算法能够得到最优的收敛速率.针对这一问题,Rakhlin 等人在没有改变 SGD 算法迭代步骤的前提下,在解决强凸随机优化问题时,仅仅将算法的平均输出方式以后半部分平均的 α-suffix 技巧来代替,最终达到了 $O(1/T)$的最优收敛速率.但是该平均技巧与标准的平均方式相比也带来了一个缺点,导致了不能按照 on-the-fly 的方式计算.2012 年,Lacoste-Julien 等人[8]提出一种加权平均方式,与 α-suffix 平均不同的是,该方法仅对 SGD 算法的每步的输出解乘以一个权重,从而在得到最优的收敛速率的同时,还保证了 on-the-fly 的计算方式.

众所周知,机器学习优化问题一般以“正则化项+损失函数”的形式存在[2],早期的学习算法将正则化项和损失函数组成的目标函数当做一个整体来考虑,并不区别看待,SGD 就属于这种黑箱方法.然而,近些年的研究成果表明:正则化项和损失函数往往有着机器学习的特殊含义,如 $L1$ 正则化保证了解的稀疏性[9];$L2$ 正则化具有比较明确支持向量的含义[10]等.著名优化领域专家 Nesterov 也曾指出:“黑箱方法在凸优化问题上的重要性将不可逆转地消失,彻底地取而代之的是巧妙运用问题结构的新算法[11]”.此后,广大的研究者们开始意识到那些能够充分发掘优化问题本身结构的算法将起到越来越重要的地位.

由于 SGD 方法无法充分挖掘优化问题的学习结构,在求解随机优化问题时,极大地掩盖了正则化项的作用.为此,众多学者都致力于搜寻可以突出机器学习结构含义的算法[12−14],其中,COMID(composite objective mirror descent)算法[13]以其简单高效而倍受学者们推崇.COMID 是 Duchi 等人对经典的镜面下降算法 MD[15,16]的突破性改进,与 MD 算法相比,COMID 在优化过程中将正则化项和损失函数区分看待,只对损失函数进行近似线性展开,从而保证了正则化项的结构.此外,该算法可以直接使用软阈值方法[12]对优化子问题解析求解,减少了计算代价.COMID 算法先以在线形式被提出,后又扩展为随机算法.但是,在求解强凸优化问题时, COMID 仅能够得到 $O(\ln T/T)$的收敛速率.

综合以上分析,能否把加权平均技巧与 COMID 算法相结合,将 COMID 算法求解强凸优化问题的收敛速率加速到 $O(1/T)$?从 COMID 算法的收敛性分析中不难发现:当正则化和损失函数分开考虑时,其证明过程中的主要困难而又复杂之处体现在如何界定正则化项交错问题导致的上界.本文在 COMID 算法中引入了类似文献[8]中的加权平均技巧,提出了一种 HRMD-W(hybrid regularized mirror descent with weighted averaging)算法,在考虑 $L1$+$L2$+Hinge 的混合正则化项的随机优化问题时,我们首先使用软阈值方法给出交错项的上界;其次,通过巧妙的选取步长,在理论上证明 HRMD-W 具有 $O(1/T)$的收敛速率,并且可以使用 on-the-fly 方法减少计算代价;最后,在大规模数据库上的实验结果表明:HRMD-W 算法在保证与 COMID 算法具有相同稀疏性的同时,获得了较高的正确率.

1 SGD,COMID 算法及加速技巧

不失一般性,我们在此仅讨论最简单的二分类问题,假设训练样本集独立同分布,并且表示为 $S=\{(x_1,y_1),\ldots,(x_m,y_m)\}$,其中,$(x_i,y_i)\in R^n\times\{-1,+1\}$,$m$ 为样本个数,n 为样本维数.S 中的样本满足独立同分布,我们使用 $\xi=(\boldsymbol{x},y)$表示随机抽取的样本.

1.1 SGD算法

SGD 以操作简单、计算便捷、理论分析完善,其主要执行流程见算法 1,假设 Φ为优化问题的目标函数,则

SGD 所要求解的优化问题如下:

$$\min_{w\in\Omega}\Phi(\boldsymbol{w}) \tag{1}$$

其中,$\Phi(\boldsymbol{w})=E_{\xi}[\Phi(\boldsymbol{w},\xi)]$,$\Omega$为 R^n 上的闭凸集合.

从公式(1)可以看出:SGD 的目标函数为期望形式,这正是随机算法的特点.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[6,17],即有 $E[\boldsymbol{g}_t]\in\partial\Phi(\boldsymbol{w}_t)$,因此,SGD 在算法形式上表现为每步迭代仅优化随机抽取的一个样本.此外,SGD 算法的一个重要评价指标——收敛速率,是指在数学期望下,SGD 输出解对应的目标函数值收敛于最优目标函数值的速率[6],数学表达式为

$$E[\Phi(\boldsymbol{w})-\Phi(\boldsymbol{w}^*)],$$

其中,$\boldsymbol{w}^*$为优化问题的最优解.

算法 1. SGD 算法.

1. Initialize $\boldsymbol{w}_1=\boldsymbol{0}$
2. **for** t=1 to T
3. Compute $\boldsymbol{g}_t\in\partial\Phi(\boldsymbol{w}_t,\xi_t)$, $\eta_t=1/\sigma t$
4. Compute $\boldsymbol{w}_{t+1}=P_{\Omega}(\boldsymbol{w}_t-\eta_t\boldsymbol{g}_t)$
5. **end for**
6. Output: $\bar{w}_T=(w_1+\ldots+w_T)/T$

其中,σ为强凸参数,η_t 为步长,$\boldsymbol{g}_t$ 表示$\Phi(\boldsymbol{w}_t,\xi)$的次梯度,$P_{\Omega}$表示在$\Omega$上的投影算子.

1.2 COMID算法

COMID 算法是 2010 年由 Duchi 等人提出,该算法将优化目标函数中的正则化项和损失函数区分对待,仅对损失函数进行线性近似展开,而不对正则化项作处理.此时,子问题就具有解析解,算法既能够快速收敛又能够保证正则化项的各种性质.特别地,当 $r(\boldsymbol{w})$为 $L1$ 正则化项时,能够得到稀疏解.具体来说,在结构学习框架下,COMID 算法的主要迭代形式如下:

$$\boldsymbol{w}_{t+1}=\arg\min_{w\in\Omega}\{\eta_t\langle\boldsymbol{g}_t,\boldsymbol{w}\rangle+\eta_t r(\boldsymbol{w})+B_{\varphi}(\boldsymbol{w},\boldsymbol{w}_t)\} \tag{2}$$

其中,$r(\boldsymbol{w})$为正则化项,若用 $l(\boldsymbol{w},\xi)$表示损失函数,则 $\boldsymbol{g}_t$ 为损失函数 $l(\boldsymbol{w}_t,\xi_t)$的次梯度,$B_{\varphi}(\boldsymbol{w},\boldsymbol{w}_t)$为 Bregman[18]. COMID 算法所要优化的目标函数为$\Phi(\boldsymbol{w})=E_{\xi}[r(\boldsymbol{w})+l(\boldsymbol{w},\xi)]$,其执行过程与 SGD 类似,区别在于 COMID 所要求解的子问题为公式(2).

在求解强凸优化问题时,当步长取 $O(1/t)$时,SGD 和 COMID 算法都只达到 $O(\log T/T)$的收敛速率,为此,Shalev-Shwartz 和 Duchi 等人给出如下定理:

定理 1[5,13]. 假设 $\boldsymbol{w}^*$为优化问题(1)的最优解,取步长$\eta_t=O(1/t)$,则当 SGD 或 COMID 算法求解强凸优化问题并运行 T 次迭代后,其收敛速率满足:

$$E[\Phi(\bar{\boldsymbol{w}}_T)]-\Phi(\boldsymbol{w}^*)\leqslant O(\log T/T),$$

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

1.3 SGD的加速技巧

众所周知,强凸优化问题的最优收敛速率为 $O(1/T)$,但标准的 SGD 算法(算法 1)仅能达到 $O(\log T/T)$.为此,文献[7]通过对算法 1 的输出简单改动,使 SGD 算法得到最优的收敛速率,即对算法输出使用α-suffix 平均技巧,将原来的标准平均方式改为

$$\bar{\boldsymbol{w}}_T^{\alpha}=\frac{\boldsymbol{w}_{(1-\alpha)T+1}+\ldots+\boldsymbol{w}_T}{\alpha T},$$

其中,$\alpha\in(0,1)$,设αT 为整数.

然而,该方法不能 on-the-fly 地进行计算.通过观察可知:标准平均方式 $\bar{\boldsymbol{w}}_T$ 可以写成 $\bar{\boldsymbol{w}}_T=(1-1/T)\bar{\boldsymbol{w}}_{T-1}+(1+T)\boldsymbol{w}_T$ 的形式,在这种形式下,算法不需要运行所有迭代步骤,在运行的过程中,仅需少量的计算代价,便可对

$\bar{\boldsymbol{w}}_T$ 进行更新,这种方式对算法的执行起着极其重要的作用.α-suffix 平均方式需要存储算法每次迭代产生的 $\boldsymbol{w}_T$,或者需要事先知道算法迭代次数,固不能 on-the-fly 地进行计算.

随后,文献[8]为了克服α-suffix 平均的缺点,使用另外一种平均方式——加权平均,同样使 SGD 算法得到 $O(1/T)$的收敛速率.加权平均不同于以往仅简单对 $\boldsymbol{w}_t$取平均的做法,在算法第 t 步迭代产生 $\boldsymbol{w}_t$的同时,为 $\boldsymbol{w}_t$乘以一个权重 t,其输出形式为

$$\bar{\boldsymbol{w}}_T^w=\frac{2}{T(T+1)}\sum_{t=1}^{T}t\boldsymbol{w}_t=(1-\rho_T)\bar{\boldsymbol{w}}_{T-1}^w+\rho_T\boldsymbol{w}_T,$$

其中,$\rho_T=2/(T+1)$.

加权平均不仅能够 on-the-fly 地进行计算,而且通过巧妙地选取步长,使得其收敛速率的理论证明更加地简洁易懂.

使用α-suffix 和加权平均技巧求解强凸优化问题时,都能使优化问题的收敛速率加速到最优,即得到 $O(1/T)$的收敛速率.为此,文献[7,8]给出如下定理:

定理 2[7,8]. 令 $\boldsymbol{w}^*$为优化问题(1)的最优解,取步长 $\eta_t=O(1/t)$,SGD 算法使用α-suffix 或加权平均技巧求解强凸优化问题,其收敛速率满足:

$$E[\Phi(\bar{\boldsymbol{w}})]-\Phi(\boldsymbol{w}^*)\leqslant O(1/T).$$

当使用α-suffix 平均技巧时,$\bar{\boldsymbol{w}}=\bar{\boldsymbol{w}}_T^{\alpha}$;使用加权平均技巧时,有 $\bar{\boldsymbol{w}}=\bar{\boldsymbol{w}}_T^w$.

本文结合 COMID 和加权平均技巧,提出 HRMD-W 算法.在求解 $L1$ 和 $L2$ 混合正则化项导致的强凸优化问题时,通过克服 HRMD-W 算法收敛性分析中的关键性问题,从理论方面将该算法的收敛速率提升至最优,即,达到 $O(1/T)$.

2 HRMD-W 算法及收敛性分析

本文主要考虑特殊的具有 $L1$ 和 $L2$ 混合正则化项的随机优化问题:

$$\min_{w}E_{\xi}[\Phi(\boldsymbol{w},\xi)]=E_{\xi}\left[\lambda\|\boldsymbol{w}\|_1+\frac{\sigma}{2}\|\boldsymbol{w}\|_2^2+l(\boldsymbol{w};\xi)\right] \tag{3}$$

其中,$l(\boldsymbol{w};\xi)=\max\{0,1-y_t\langle\boldsymbol{w};\boldsymbol{x}_t\rangle\}$为样本$(\boldsymbol{x}_t,y_t)$的 Hinge 损失.不难看出:$L2$ 正则化项的引入,使整个目标函数具备了强凸性质.同时,为了保证目标函数的结构,将此强凸项并入原正则化项中,此时,正则化项为

$$r(\boldsymbol{w})=\lambda\|\boldsymbol{w}\|_1+\frac{\sigma}{2}\|\boldsymbol{w}\|_2^2.$$

易知,$r(\boldsymbol{w})$为强凸函数,满足σ-强凸[19].

算法 2. HRMD-W 算法.

1: Input parameters λ, σ. Initialize $\boldsymbol{w}_1=\boldsymbol{0}$
2: **for** t=1 to T
3: Compute $\boldsymbol{g}_t\in\partial l(\boldsymbol{w}_t,\xi_t)$, $\eta_t=2/\sigma t$
4: Compute $\boldsymbol{w}_{t+1}$ via Eq.(4)
5: **end for**
6: Output: $\bar{\boldsymbol{w}}_T^w=\frac{2}{T(T+3)}\sum_{t=1}^{T}(t+1)\boldsymbol{w}_t$

算法 2 为 HRMD-W 算法的主要执行过程,在算法中,我们取 $B_{\varphi}(\boldsymbol{w},\boldsymbol{w}_t)=\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_t\|_2^2$. 与文献[8]不同的是,算法 2 在第 t 步迭代产生 $\boldsymbol{w}_t$的同时,为 $\boldsymbol{w}_t$乘以一个权重 t+1,其输出可以写为

$$\bar{\boldsymbol{w}}_T^w=\frac{2}{T(T+3)}\sum_{t=1}^{T}(t+1)\boldsymbol{w}_t=(1-\rho_T)\bar{\boldsymbol{w}}_{T-1}^w+\rho_T\boldsymbol{w}_T,$$

其中, $\rho_T=\frac{2(T+1)}{T(T+3)}$. 由此可见, $\bar{\boldsymbol{w}}_T^w$ 可通过 on-the-fly 方式进行计算.HRMD-W 算法主要迭代步骤为

$$\begin{aligned}\boldsymbol{w}_{t+1}&=\arg\min_{w\in\Omega}\left\{\eta_t\langle\boldsymbol{g}_t,\boldsymbol{w}\rangle+\eta_t r(\boldsymbol{w})+\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_t\|^2\right\}\\&=\arg\min_{w\in\Omega}\left\{\frac{\sigma\eta_t+1}{2}\boldsymbol{w}^2+\langle\eta_t\boldsymbol{g}_t-\boldsymbol{w}_t,\boldsymbol{w}\rangle+\lambda\eta_t\|\boldsymbol{w}\|_1\right\}\end{aligned}\tag{4}$$

本文的收敛速率证明与文献[8]基本类似,但 SGD 和 COMID 算法存在着本质的区别.为此,我们给出了正则化交错项需满足的关系式,见引理 1.

引理 1. 假设 $E[\|\boldsymbol{g}_t\|]\leqslant G,\forall t,\|\boldsymbol{w}\|_\infty=\max_i|\boldsymbol{w}_i|\leqslant M$, 对公式(4)采用软阈值方法进行求解,则正则化交错项有如下关系式成立:

$$E[r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})]\leqslant A\eta_t^2+B\eta_t,$$

其中,

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

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

根据以上引理,并使用加权平均特有的证明技巧,我们很容易得出如下两个定理,附录 2 给出了详细的证明:

定理 3. 令 $\boldsymbol{w}^*$为优化问题(3)的最优解,则 HRMD-W 算法运行 T 次迭代后,其收敛速率满足如下关系:

$$E[\Phi(\bar{\boldsymbol{w}}_T^w)]-\Phi(\boldsymbol{w}^*)\leqslant\frac{(4B+2G^2)/\sigma}{T+3}+\frac{(4B+2G^2)/\sigma+8A/\sigma^2}{T(T+3)}\ln T+\frac{\sigma\|\boldsymbol{w}^*\|^2+24A/\sigma^2+(4B+2G^2)/\sigma}{T(T+3)},$$

其中,A 和 B 同引理 1.

定理 4. 令 $\boldsymbol{w}^*$为优化问题(3)的最优解,则 HRMD-W 算法输出的任意瞬时解与最优解之间满足如下关系:

$$E[\|\boldsymbol{w}_{T+1}-\boldsymbol{w}^*\|^2]\leqslant\frac{2/\sigma}{(T+1)(T+2)}\left[\frac{4B+2G^2}{\sigma}T+\left(\frac{4B+2G^2}{\sigma}+\frac{8A}{\sigma^2}\right)\ln T+\sigma\|\boldsymbol{w}^*\|^2+\frac{24A}{\sigma^2}+\frac{(4B+2G^2)}{\sigma}\right],$$

其中,A 和 B 同引理 1.

从定理 3 不难看出:不等式右边的形式为 $O(1/T)+O(\ln T/T^2)+O(1/T^2)$,其中起主导作用的项为 $O(1/T)$.所以,当 HRMD-W 算法运行 T 步后的收敛速率为最优的 $O(1/T)$.类似地,定理 4 描述了算法迭代过程的解与最优解之间欧式距离的界小于 $O(1/T)$.

3 数值实验

本节通过 4 个大规模数据库验证 HRMD-W 算法的实际效果.实验环境为 Sun Ultra45 工作站(1.6GHz UltraSPARC IIIi 处理器,4GB 内存,Sorlaris10 操作系统),实验所采用的比较算法均在 LIBLINEAR[20]平台上实现.

3.1 实验数据库和比较算法

实验所采用的 4 个大规模数据库分别为 CCAT,astro-physic,a9a 和 covtype,表 1 给出了这 4 个数据库的详细描述.

Table 1 Description of datasets

表 1 实验数据库描述

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

实验对 3 种算法进行比较,分别为 HRMD-W 算法、SGD-W 算法[8]和采用 $L1$+$L2$ 混合正则化的 COMID 算

法,记为 HRCOMID 算法.对于使用 $L2$ 正则化项的强凸优化问题,SGD-W 算法得到了 $O(1/T)$的最优收敛速率;根据文献[13]易知,HRCOMID 算法的收敛速率为 $O(\log T/T)$.

3.2 实验方法及结论

实验过程中,我们对每个数据库的样本采取随机抽取的方式,算法进行 10 000 次迭代后终止.为公平起见,算法中的参数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,最终结果为 10 次结果的平均值及均方差.

图 1 为 3 种算法的收敛速率比较图,其中,横坐标表示迭代次数;纵坐标表示当前目标函数值与最优目标函数值之差,在此用 $\Phi(\bar{\boldsymbol{w}})-\Phi(\boldsymbol{w}^*)$ 表示,具体计算时,最优目标函数值取迭代中最小的目标函数值.

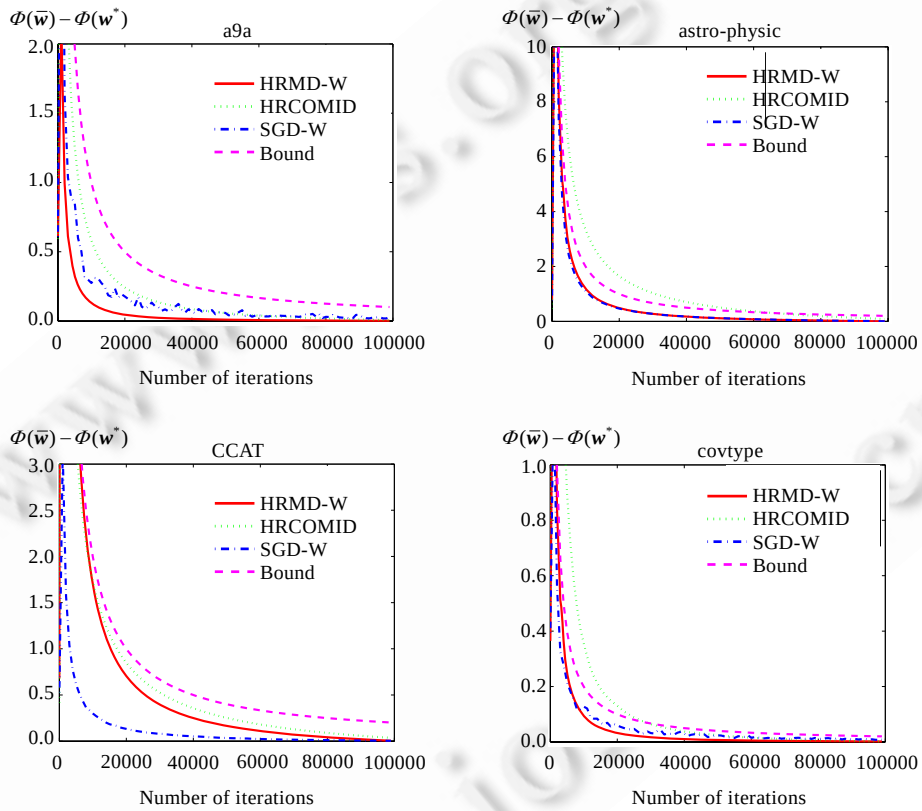

Fig.1 Comparisons of convergence rate

图 1 目标函数收敛速率比较图

从图 1 可以看出:HRMD-W 算法在 astro-physic 和 covtype 数据库上与 SGD-W 算法收敛速率相当;在 a9a 数据库的收敛速率快于 SGD-W;但在 CCAT 数据库慢于 SGD-W 算法.总的来说,HRMD-W 基本达到与 SGD-W 算法相同的收敛速率,且均快于 HRCOMID 算法.图中代表 Bound 的双划线表示定理 2 所给出的算法收敛速率的界.可以看出,HRMD-W 算法的实际收敛速率曲线均在其相应界的下方,进而说明本文理论分析的正确性.

表 2 给出算法在 4 个数据库上的测试错误率及方差.比较表中 3 种算法很容易看出:算法 HRMD-W 的测试错误率最小,准确率最高.众所周知:方差在一定程度上反映算法的稳定性,方差越小,表示算法的稳定性越好.从表 2 可以看出:HRMD-W 和 SGD-W 算法的稳定性优于 HRCOMID 算法,这是因为在算法输出解中使用加权技巧,从而使得算法的稳定性增加.

Table 2 Comparisons of test error and variance

表 2 测试错误率和方差比较

算法/数据库	a9a	CCAT	astro-physic	covtype
HRMD-W	0.1534±0.0008	0.0783±0.0007	0.0402±0.0006	0.2323±0.0014
HRCOMID	0.1570±0.0014	0.0932±0.0020	0.0526±0.0025	0.2357±0.0021
SGD-W	0.1534±0.0013	0.0865±0.0010	0.0405±0.0006	0.2357±0.0014

图 2 为 3 种算法的稀疏度比较图.稀疏度是指算法输出解向量中零维所占的比例,解的稀疏度越高,表示算法稀疏性越好.比较 HRMD-W 和 SGD-W 算法容易发现:在 a9a 和 astro-physic 数据库上,两者的稀疏性相差不大;在 CCAT 和 covtype 上,HRMD-W 算法的稀疏性较好.比较 HRMD-W 和 HRCOMID,从图中可以看出:对于 covtype 数据库,HRCOMID 算法的稀疏性稍好一些;在 a9a,两种算法稀疏性相当;但在 astro-physic 和 CCAT 数据库上,HRMD-W 的稀疏性明显好于 HRCOMID 算法.

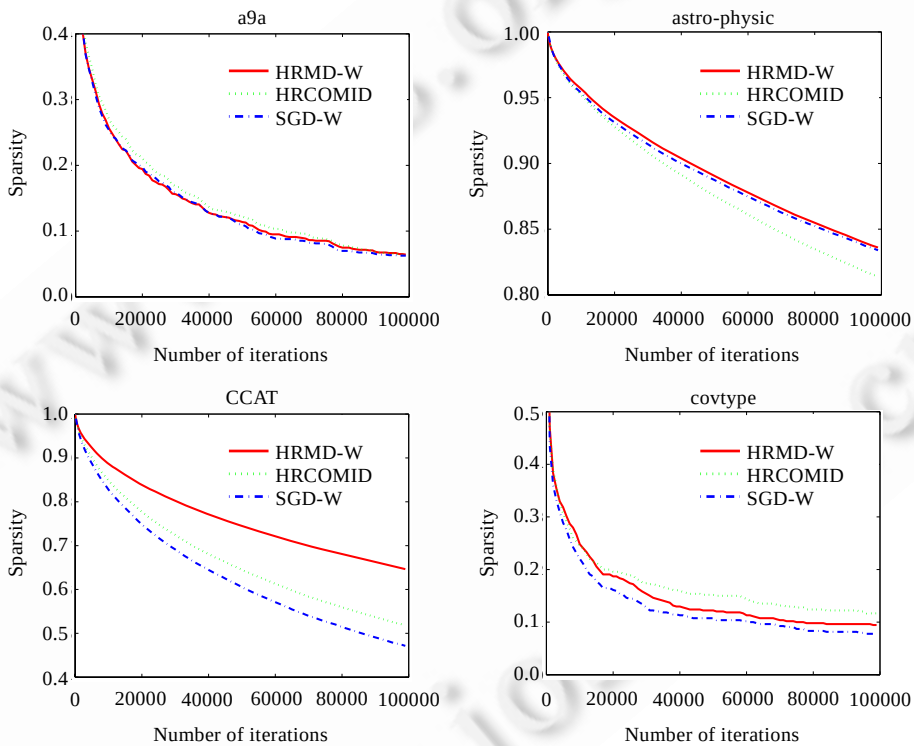

Fig.2 Comparisons sparsity

图 2 稀疏度比较图

4 总结与展望

本文在混合正则化项的镜面下降算法中引入加权技巧,提出了 HRMD-W 算法.该算法在保证正确率的前提下,以 on-the-fly 的计算方式减少计算代价,并且在理论上证明算法达到 $O(1/T)$的最优收敛速率.最后,通过实验对算法的性能进行验证,说明理论分析的正确性和所提算法的有效性.

对强凸优化问题,通过使用加权平均技巧,不仅能得到最优的收敛速率,而且在一定程度上也能减少算法的计算代价和提高算法的稳定性.这仅仅是将该技巧用于算法的最终输出上,近年也出现了将该技巧用于梯度上[21].能否将加权平均用于算法的迭代过程中,是我们下一步努力的方向.

References:

[1] 孙正雅,陶卿.统计机器学习综述:损失函数与优化求解.中国计算机学会通讯,2009,5(8):7−14.

[2] Tao Q, Gao QK, Jiang JY, Chu DJ. Survey of solving the optimization problems for sparse learning. Ruan Jian Xue Bao/Journal of Software, 2013,24(11):2498−2507 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4479.htm [doi: 10.3724/SP.J.1001.2013.044790]

[3] Robbins H, Monro S. A stochastic approximation method. The Annuals of Mathematical Statistics, 1951,22:400−407. [doi: 10.1214/aoms/1177729586]

[4] Kiefer J, Wolfowitz J. Stochastic estimation of the maximum of a regression function. The Annuals of Mathematical Statistics, 1952,23:462−466.

[5] Shalev-Shwartz S, Singer Y, Srebro N. Pegasos: Primal estimated sub-gradient solver for SVM. In: Proc. of the 24th Int'l Conf. on Machine Learning. ACM Press, 2007. 807−814.

[6] Hazan E, Kale S. Beyond the regret minimization barrier: An optimal algorithm for stochastic strongly-convex optimization. Journal of Machine Learning Research-Proceedings Track, 2011,19:421−436.

[7] Rakhlin A, Shamir O, Sridharan K. Making gradient descent optimal for strongly convex stochastic optimization. In: Proc. of the 29th Int'l Conf. on Machine Learning. 2012. 449−456. http://arxiv.org/abs/1109.5647

[8] Lacoste-Julien S, Schmidt M, Bach F. A simpler approach to obtaining an $o(1/t)$ convergence rate for projected stochastic subgradient descent. arXiv preprint arXiv:1212.2002, 2012.

[9] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of Royal Statistical Society, Series B, 1996,58(1):267−288.

[10] Shalev-Shwartz S, Zhang T. Stochastic dual coordinate ascent methods for regularized loss minimization. Journal of Machine Learning Research, 2013,14:567−599.

[11] Nesterov Y. How to advance in structural convex optimization. OPTIMA: Mathematical Programming Society Newsletter, 2008,78:2−5.

[12] Xiao L. Dual averaging methods for regularized stochastic learning and online optimization. The Journal of Machine Learning Research, 2010,11:2543−2596.

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

[14] Ouyang H, He N, Tran L, Gray A. Stochastic alternating direction method of multipliers. In: Proc. of the 30th Int'l Conf. on Machine Learning. 2013. 80−88. http://jmlr.org/proceedings/papers/v28/ouyang13.html

[15] Nemirovski A, Yudin D. Problem Complexity and Method Efficiency in Optimization. Wiley-Interscience Series in Discrete Mathematics, John Wiley-Interscience Publication, 1983.

[16] Beck A, Teboulle M. Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters, 2003,31(3):167−175. [doi: 10.1016/S0167-6377(02)00231-6]

[17] Nemirovski A, Juditsky A, Lan G, Shapiro A. Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization, 2009,19(4):1574−1609. [doi: 10.1137/070704277]

[18] Bregman LM. The relaxation method of finding the common point convex sets and its application to the solution of problems in convex programming. USSR Computational Mathematics and Mathematical Physics, 1967,7:200−217. [doi: 10.1016/0041-5553(67)90040-7]

[19] Bertsekas D. Convex Analysis and Optimization. Athena: Scientific, 2003.

[20] Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A library for large linear classification. The Journal of Machine Learning Research, 2008,9:1871−1874.

[21] Chen X, Lin Q, Pena J. Optimal regularized dual averaging methods for stochastic optimization. In: Proc.of the Advances in Neural Information Processing Systems. 2012. 404−412. http://papers.nips.cc/paper/4543-optimal-regularized-dual-averaging-methods-for-stochastic-optimization.pdf

附中文参考文献:

[2] 陶卿,高乾坤,姜纪远,储德军.稀疏学习优化问题的求解综述.软件学报,2013,24(11):2498−2507. http://www.jos.org.cn/1000-9825/4479.htm [doi: 10.3724/SP. J.1001.2013.044790]

附录 1. 引理 1 的证明

证明:公式(4)所要解决的优化问题为

$$\boldsymbol{w}_{t+1}=\arg\min_{\boldsymbol{w}\in\Omega}\left\{\frac{\sigma\eta_t+1}{2}\boldsymbol{w}^2+\langle\eta_t\boldsymbol{g}_t-\boldsymbol{w}_t,\boldsymbol{w}\rangle+\lambda\eta_t\|\boldsymbol{w}\|_1\right\}.$$

由于向量 $\boldsymbol{w}$ 的每一维不相关,因此可以将其转换为如下单维优化问题:

$$\boldsymbol{w}_{t+1,j}=\arg\min_{\boldsymbol{w}_j}\left\{\frac{\sigma\eta_t+1}{2}\boldsymbol{w}_j^2+\langle\eta_t\boldsymbol{g}_{t,j}-\boldsymbol{w}_{t,j},\boldsymbol{w}_j\rangle+\lambda\eta_t|\boldsymbol{w}_j|\right\},1\leqslant j\leqslant n.$$

满足二次项+一次项+正则化项条件,使用软阈值方法解得:

$$\boldsymbol{w}_{t+1,j}=\begin{cases}0, & \text{if } |\eta_t\boldsymbol{g}_{t,j}-\boldsymbol{w}_{t,j}|\leqslant\lambda\eta_t\\ \dfrac{1}{\sigma\eta_t+1}[\boldsymbol{w}_{t,j}-\eta_t\boldsymbol{g}_{t,j}-\lambda\eta_t\operatorname{sgn}(\boldsymbol{w}_{t,j}-\eta_t\boldsymbol{g}_{t,j})], & \text{otherwise}\end{cases}.$$

分两种情况讨论:

(1) 当$|\boldsymbol{w}_{t,j}-\eta_t\boldsymbol{g}_{t,j}|\leqslant\lambda\eta_t$,即$|\boldsymbol{w}_{t,j}|-|\eta_t\boldsymbol{g}_{t,j}|\leqslant\lambda\eta_t$时,有 $\boldsymbol{w}_{t+1,j}=0$.此时,单维的正则化交错项为

$$\begin{aligned}r(\boldsymbol{w}_t)_j-r(\boldsymbol{w}_{t+1})_j&=r(\boldsymbol{w}_t)_j\\&=\lambda|\boldsymbol{w}_{t,j}|+\frac{\sigma}{2}\boldsymbol{w}_{t,j}^2\\&\leqslant\lambda\eta_t(|\boldsymbol{g}_{t,j}|+\lambda)+\frac{\sigma\eta_t^2}{2}(|\boldsymbol{g}_{t,j}|+\lambda)^2\\&=\frac{\sigma\eta_t^2}{2}|\boldsymbol{g}_{t,j}|^2+\lambda\eta_t(1+\sigma\eta_t)|\boldsymbol{g}_{t,j}|+\lambda^2\eta_t\left(1+\frac{\sigma\eta_t}{2}\right).\end{aligned}$$

所以,该情况正则化项满足:

$$\begin{aligned}E[r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})]&=\sum_{j=1}^{n}E[r(\boldsymbol{w}_t)_j-r(\boldsymbol{w}_{t+1})_j]\\&=\sum_{j=1}^{n}E\left[\frac{\sigma\eta_t^2}{2}|\boldsymbol{g}_{t,j}|^2+\lambda\eta_t(1+\sigma\eta_t)|\boldsymbol{g}_{t,j}|+\lambda^2\eta_t\left(1+\frac{\sigma\eta_t}{2}\right)\right]\\&=E\left[\frac{\sigma\eta_t^2}{2}\|\boldsymbol{g}_t\|_2^2+\lambda\eta_t(1+\sigma\eta_t)\|\boldsymbol{g}_t\|_1+n\lambda^2\eta_t\left(1+\frac{\sigma\eta_t}{2}\right)\right]\\&\leqslant E\left[\frac{\sigma\eta_t^2}{2}\|\boldsymbol{g}_t\|_2^2+\lambda\sqrt{n}\eta_t(1+\sigma\eta_t)\|\boldsymbol{g}_t\|_2+n\lambda^2\eta_t\left(1+\frac{\sigma\eta_t}{2}\right)\right]\\&\leqslant\frac{\sigma\eta_t^2}{2}G^2+\lambda\sqrt{n}\eta_t(1+\sigma\eta_t)G+n\lambda^2\eta_t\left(1+\frac{\sigma\eta_t}{2}\right)\\&=\frac{\sigma}{2}(G^2+2\sqrt{n}\lambda G+n\lambda^2)\eta_t^2+\lambda(\sqrt{n}G+n\lambda)\eta_t.\end{aligned}$$

为方便整理,我们定义两个参数来简化证明,表示如下:

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

$$B_1=\lambda(\sqrt{n}G+n\lambda).$$

得到:

$$E[r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})]\leqslant A_1\eta_t^2+B_1\eta_t.$$

(2) 当$|\boldsymbol{w}_{t,j}-\eta_t\boldsymbol{g}_{t,j}|>\lambda\eta_t$时,有:

$$\boldsymbol{w}_{t+1,j}=\frac{1}{\sigma\eta_t+1}(\boldsymbol{w}_{t,j}-\eta_t\boldsymbol{g}_{t,j}-\lambda\eta_t\operatorname{sgn}(\boldsymbol{w}_{t,j}-\eta_t\boldsymbol{g}_{t,j})).$$

为方便证明,用 sgn 表示 $\operatorname{sgn}(\boldsymbol{w}_{t,j}-\eta_t\boldsymbol{g}_{t,j})$.此时:

$$
\begin{aligned}
r(\boldsymbol{w}_t)_j - r(\boldsymbol{w}_{t+1})_j &= r(\boldsymbol{w}_t)_j \\
&= \lambda|\boldsymbol{w}_{t,j}| - \lambda|\boldsymbol{w}_{t+1,j}| + \frac{\sigma}{2}\boldsymbol{w}_{t,j}^2 - \frac{\sigma}{2}\boldsymbol{w}_{t+1,j}^2 \\
&\leqslant \lambda\left|\boldsymbol{w}_{t,j} - \frac{1}{\sigma\eta_t+1}(\boldsymbol{w}_{t,j} - \eta_t\boldsymbol{g}_{t,j} - \lambda\eta_t\operatorname{sgn})\right| + \frac{\sigma}{2}\left[|\boldsymbol{w}_{t,j}|^2 - \frac{1}{(\sigma\eta_t+1)^2}|\boldsymbol{w}_{t,j} - \eta_t\boldsymbol{g}_{t,j} - \lambda\eta_t\operatorname{sgn}|^2\right] \\
&\leqslant \frac{\lambda\eta_t}{\sigma\eta_t+1}(\sigma|\boldsymbol{w}_{t,j}| + |\boldsymbol{g}_{t,j}| + \lambda) + \frac{\sigma}{2(\sigma\eta_t+1)^2}[(\sigma\eta_t+1)^2|\boldsymbol{w}_{t,j}|^2 - (|\boldsymbol{w}_{t,j}| - \eta_t|\boldsymbol{g}_{t,j} + \lambda\operatorname{sgn}|)^2] \\
&\leqslant \frac{\lambda\eta_t}{\sigma\eta_t+1}(\sigma|\boldsymbol{w}_{t,j}| + |\boldsymbol{g}_{t,j}| + \lambda) + \\
&\quad \frac{\sigma\eta_t}{2(\sigma\eta_t+1)^2}[\sigma(\sigma\eta_t+2)|\boldsymbol{w}_{t,j}|^2 + 2|\boldsymbol{w}_{t,j}|(|\boldsymbol{g}_{t,j}| + \lambda) - \eta_t|\boldsymbol{g}_{t,j} + \lambda\operatorname{sgn}|^2] \\
&\leqslant \frac{\lambda\eta_t}{\sigma\eta_t+1}(\sigma|\boldsymbol{w}_{t,j}| + |\boldsymbol{g}_{t,j}| + \lambda) + \frac{\sigma\eta_t}{2(\sigma\eta_t+1)^2}[\sigma(\sigma\eta_t+2)|\boldsymbol{w}_{t,j}|^2 + 2|\boldsymbol{w}_{t,j}|(|\boldsymbol{g}_{t,j}| + \lambda) - \eta_t(|\boldsymbol{g}_{t,j}| - \lambda)^2] \\
&\leqslant \frac{\lambda\eta_t}{\sigma\eta_t+1}(\sigma M + |\boldsymbol{g}_{t,j}| + \lambda) + \frac{\sigma\eta_t}{2(\sigma\eta_t+1)^2}[\sigma(\sigma\eta_t+2)M^2 + 2M(|\boldsymbol{g}_{t,j}| + \lambda) - \eta_t(|\boldsymbol{g}_{t,j}| - \lambda)^2] \\
&= \frac{\eta_t}{(\sigma\eta_t+1)^2}\left[-\frac{\sigma\eta_t}{2}|\boldsymbol{g}_{t,j}|^2 + (2\lambda\sigma\eta_t + \lambda + \sigma M)|\boldsymbol{g}_{t,j}| + \right. \\
&\qquad \left. \frac{\sigma^3M^2 + 2\lambda\sigma^2M + \lambda^2\sigma}{2}\eta_t + \sigma^2M^2 + 2\lambda\sigma M + \lambda^2\right] \\
&\leqslant \eta_t\left[-\frac{\sigma\eta_t}{2}|\boldsymbol{g}_{t,j}|^2 + (2\lambda\sigma\eta_t + \lambda + \sigma M)|\boldsymbol{g}_{t,j}| + \frac{\sigma^3M^2 + 2\lambda\sigma^2M + \lambda^2\sigma}{2}\eta_t + \sigma^2M^2 + 2\lambda\sigma M + \lambda^2\right].
\end{aligned}
$$

所以,该情况正则化项满足:

$$
\begin{aligned}
E[r(\boldsymbol{w}_t) - r(\boldsymbol{w}_{t+1})] &= \sum_{j=1}^{n} E[r(\boldsymbol{w}_t)_j - r(\boldsymbol{w}_{t+1})_j] \\
&= \sum_{j=1}^{n}\eta_t E\left[-\frac{\sigma\eta_t}{2}|\boldsymbol{g}_{t,j}|^2 + (2\lambda\sigma\eta_t + \lambda + \sigma M)|\boldsymbol{g}_{t,j}| + \frac{\sigma^3M^2 + 2\lambda\sigma^2M + \lambda^2\sigma}{2}\eta_t + \sigma^2M^2 + 2\lambda\sigma M + \lambda^2\right] \\
&= \eta_t E\left[-\frac{\sigma\eta_t}{2}\|\boldsymbol{g}\|_2^2 + (2\lambda\sigma\eta_t + \lambda + \sigma M)\|\boldsymbol{g}\|_1 + \right. \\
&\qquad \left. \frac{n(\sigma^3M^2 + 2\lambda\sigma^2M + \lambda^2\sigma)}{2}\eta_t + n(\sigma^2M^2 + 2\lambda\sigma M + \lambda^2)\right] \\
&\leqslant \eta_t E\left[-\frac{\sigma\eta_t}{2}\|\boldsymbol{g}\|_2^2 + (2\lambda\sigma\eta_t + \lambda + \sigma M)\|\boldsymbol{g}\|_1 + \right. \\
&\qquad \left. \frac{n(\sigma^3M^2 + 2\lambda\sigma^2M + \lambda^2\sigma)}{2}\eta_t + n(\sigma^2M^2 + 2\lambda\sigma M + \lambda^2)\right] \\
&\leqslant \eta_t\left[-\frac{\sigma\eta_t}{2}G^2 + (2\lambda\sigma\eta_t + \lambda + \sigma M)\sqrt{n}G + \frac{n(\sigma^3M^2 + 2\lambda\sigma^2M + \lambda^2\sigma)}{2}\eta_t + n(\sigma^2M^2 + 2\lambda\sigma M + \lambda^2)\right] \\
&= \sigma\left[\frac{1}{2}n(\sigma^2M^2 + 2\lambda\sigma M + \lambda^2) - \frac{G^2}{2} + 2\lambda\sqrt{n}G\right]\eta_t^2 + [(\lambda + \sigma M)\sqrt{n}G + n(\sigma^2M^2 + 2\lambda\sigma M + \lambda^2)]\eta_t.
\end{aligned}
$$

同样地,定义两个参数来简化证明,表示如下:

令:

$$
\begin{aligned}
A_2 &= \sigma\left(\frac{n(\sigma^2M^2 + 2\lambda\sigma M + \lambda^2)}{2} - \frac{G^2}{2} + 2\lambda\sqrt{n}G\right), \\
B_2 &= (\lambda + \sigma M)\sqrt{n}G + n(\sigma^2M^2 + 2\lambda\sigma M + \lambda^2),
\end{aligned}
$$

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

综上两种情况,我们令:

$$A=\max\{A_1,A_2\},$$
$$B=\max\{B_1,B_2\}=B_2,$$

可得: $r(\boldsymbol{w}_t)-r(\boldsymbol{w}_{t+1})\leqslant A\eta_t^2+B\eta_t$. □

附录 2. 定理 3 和定理 4 的证明

证明:根据 MRMD-W 算法的主要迭代步骤:

$$\boldsymbol{w}_{t+1}=\arg\min_{\boldsymbol{w}\in\Omega}\left\{\eta_t\langle\boldsymbol{g}_t,\boldsymbol{w}\rangle+\eta_t r(\boldsymbol{w})+\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_t\|^2\right\},$$

其中,$\boldsymbol{g}_t=\partial f(\boldsymbol{w}_t,\xi_t)$.

设 $r'(\boldsymbol{w}_{t+1})\in\partial r(\boldsymbol{w}_{t+1})$,由约束优化问题的一阶最优性条件得:

$$\langle\boldsymbol{w}-\boldsymbol{w}_{t+1},\boldsymbol{w}_{t+1}-\boldsymbol{w}_t+\eta_t\boldsymbol{g}_t+\eta_t r'(\boldsymbol{w}_{t+1})\rangle\geqslant 0.$$

特别地,当 $\boldsymbol{w}=\boldsymbol{w}^*$时:

$$\langle\boldsymbol{w}^*-\boldsymbol{w}_{t+1},\boldsymbol{w}_{t+1}-\boldsymbol{w}_t+\eta_t\boldsymbol{g}_t+\eta_t r'(\boldsymbol{w}_{t+1})\rangle\geqslant 0.$$

进一步化简得:

$$\langle\boldsymbol{w}^*-\boldsymbol{w}_{t+1},\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\rangle\geqslant\eta_t\langle\boldsymbol{w}_{t+1}-\boldsymbol{w}^*,\boldsymbol{g}_t+r'(\boldsymbol{w}_{t+1})\rangle.$$

又因为 $f(\boldsymbol{w},\xi)$为一般凸函数,$r(\boldsymbol{w})$函数满足σ-强凸,所以:

$$\begin{aligned}
f(\boldsymbol{w}_t,\xi)+r(\boldsymbol{w}_{t+1})-f(\boldsymbol{w}^*,\xi)-r(\boldsymbol{w}^*)&\leqslant\boldsymbol{g}_t(\boldsymbol{w}_t-\boldsymbol{w}^*)+r'(\boldsymbol{w}_{t+1})(\boldsymbol{w}_{t+1}-\boldsymbol{w}^*)-\frac{\sigma}{2}\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2\\
&=\langle\boldsymbol{w}_{t+1}-\boldsymbol{w}^*,\boldsymbol{g}_t+r'(\boldsymbol{w}_{t+1})\rangle+\langle\boldsymbol{g}_t,\boldsymbol{w}_t-\boldsymbol{w}_{t+1}\rangle-\frac{\sigma}{2}\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2\\
&\leqslant\frac{1}{\eta_t}\langle\boldsymbol{w}^*-\boldsymbol{w}_{t+1},\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\rangle+\langle\sqrt{\eta_t}\boldsymbol{g}_t,\sqrt{1/\eta_t}(\boldsymbol{w}_t-\boldsymbol{w}_{t+1})\rangle-\frac{\sigma}{2}\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2\\
&\leqslant\frac{1}{2\eta_t}(\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2-\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|^2-\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2)+\\
&\quad\frac{1}{2\eta_t}\|\boldsymbol{w}_{t+1}-\boldsymbol{w}_t\|^2+\frac{\eta_t}{2}\|\boldsymbol{g}_t\|^2-\frac{\sigma}{2}\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2\\
&\leqslant\frac{1}{2\eta_t}(\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2-\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2)+\frac{\eta_t}{2}\|\boldsymbol{g}_t\|^2-\frac{\sigma}{2}\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2.
\end{aligned}$$

所以有:

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

即:

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

对上式两边取期望得:

$$E[\Phi(\boldsymbol{w}_t)-\Phi(\boldsymbol{w}^*)]\leqslant A\eta_t^2+\left(B+\frac{G^2}{2}\right)\eta_t+\frac{1}{2\eta_t}(E[\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2]-E[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2])-\frac{\sigma}{2}E[\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2].$$

取 $\eta_t=2/\sigma t$,上式得:

$$E[\Phi(\boldsymbol{w}_t)-\Phi(\boldsymbol{w}^*)]\leqslant\frac{4A}{\sigma^2}\cdot\frac{1}{t^2}+\frac{2B+G^2}{\sigma}\cdot\frac{1}{t}+\frac{\sigma}{4}E[t\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2-(t+2)\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2].$$

对上式两边同乘以$(t+1)$,并从 $t=1$ 到 T 求和得:

$$\sum_{t=1}^{T}(t+1)E[\Phi(\boldsymbol{w}_t)-\Phi(\boldsymbol{w}^*)]\leqslant\frac{4A}{\sigma^2}\sum_{t=1}^{T}\left(\frac{1}{t}+\frac{1}{t^2}\right)+\frac{2B+G^2}{\sigma}\sum_{t=1}^{T}\left(\frac{1}{t}+1\right)+\frac{\sigma}{4}\sum_{t=1}^{T}E[t(t+1)\|\boldsymbol{w}_t-\boldsymbol{w}^*\|^2-(t+1)(t+2)\|\boldsymbol{w}_{t+1}-\boldsymbol{w}^*\|^2].$$

因为,

$$\sum_{t=1}^{T}\frac{1}{t}\leqslant\ln T+1,$$

$$\sum_{t=1}^{T}\frac{1}{t^2}\leqslant 2-\frac{1}{T},$$

$$\sum_{t=1}^{T}(t+1)E(\Phi(\boldsymbol{w}_t))-\frac{T(T+3)}{2}\Phi(\boldsymbol{w}^*)\leqslant\frac{4A}{\sigma^2}\left(\ln T+3-\frac{1}{T}\right)+\frac{2B+G^2}{\sigma}(T+\ln T+1)+\frac{\sigma}{4}E[2\|\boldsymbol{w}_1-\boldsymbol{w}^*\|^2-(T+1)(T+2)\|\boldsymbol{w}_{T+1}-\boldsymbol{w}^*\|^2],$$

$$\frac{2}{T(T+3)}\sum_{t=1}^{T}(t+1)E(\Phi(\boldsymbol{w}_t))-\Phi(\boldsymbol{w}^*)\leqslant\frac{8A/\sigma^2}{T(T+3)}(\ln T+3)+\frac{(4B+2G^2)C/\sigma}{T(T+3)}(T+\ln T+1)+\frac{\sigma/2}{T(T+3)}E[2\|\boldsymbol{w}_1-\boldsymbol{w}^*\|^2-(T+1)(T+2)\|\boldsymbol{w}_{T+1}-\boldsymbol{w}^*\|^2],$$

$$0\leqslant E\left[\Phi\left(\frac{2}{T(T+3)}\sum_{t=1}^{T}(t+1)\boldsymbol{w}_t\right)\right]-\Phi(\boldsymbol{w}^*)\leqslant\frac{8A/\sigma^2}{T(T+3)}(\ln T+3)+\frac{(4B+2G^2)/\sigma}{T(T+3)}(T+\ln T+1)+\frac{\sigma}{T(T+3)}E[\|\boldsymbol{w}^*\|^2]-\frac{(T+1)(T+2)}{T(T+3)}\cdot\frac{\sigma}{2}E[\|\boldsymbol{w}_{T+1}-\boldsymbol{w}^*\|^2],$$

所以有: $E[\|\boldsymbol{w}_{T+1}-\boldsymbol{w}^*\|^2]\leqslant\frac{2/\sigma}{(T+1)(T+2)}\left[\frac{4B+2G^2}{\sigma}T+\left(\frac{4B+2G^2}{\sigma}+\frac{8A}{\sigma^2}\right)\ln T+\sigma\|\boldsymbol{w}^*\|^2+\frac{24A}{\sigma^2}+\frac{(4B+2G^2)}{\sigma}\right]$.

令 $\overline{\boldsymbol{w}}_T^w=\frac{2}{T(T+3)}\sum_{t=1}^{T}(t+1)\boldsymbol{w}_t$,所以有:

$$E[\Phi(\overline{\boldsymbol{w}}_T^w)]-\Phi(\boldsymbol{w}^*)\leqslant\frac{(4B+2G^2)/\sigma}{T+3}+\frac{(4B+2G^2)/\sigma+8A/\sigma^2}{T(T+3)}\ln T+\frac{\sigma\|\boldsymbol{w}^*\|^2+24A/\sigma^2+(4B+2G^2)/\sigma}{T(T+3)}.\quad\square$$

邵言剑(1990－),男,江苏镇江人,硕士,主要研究领域为凸优化及其在机器学习中的应用.

E-mail: shy.jian@gmail.com

姜纪远(1989－),男,硕士,主要研究领域为机器学习,模式识别.

E-mail: jyjianggle@gmail.com

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

E-mail: taoqing@gmail.com

周柏(1984－),男,硕士,主要研究领域为模式识别,人工智能.

E-mail: baimoon1984@gmail.com