ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 1,Jan. 2017,pp. 17－25
DOI:10.16337/j.1004-9037.2017.01.002

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

机器学习随机优化方法的个体收敛性研究综述

陶　卿[1]　马　坡[1]　张梦晗[1]　陶　蔚[2]

(1.中国人民解放军陆军军官学院十一系,合肥,230031; 2.解放军理工大学指挥信息系统学院,南京,210007)

摘　要:随机优化方法是求解大规模机器学习问题的主流方法,其研究的焦点问题是算法是否达到最优收敛速率与能否保证学习问题的结构。目前,正则化损失函数问题已得到了众多形式的随机优化算法,但绝大多数只是对迭代进行平均的输出方式讨论了收敛速率,甚至无法保证最为典型的稀疏结构。与之不同的是,个体解能很好保持稀疏性,其最优收敛速率已经作为 open 问题被广泛探索。另外,随机优化普遍采用的梯度无偏假设往往不成立,加速方法收敛界中的偏差在有偏情形下会随迭代累积,从而无法应用。本文对一阶随机梯度方法的研究现状及存在的问题进行综述,其中包括个体收敛速率、梯度有偏情形以及非凸优化问题,并在此基础上指出了一些值得研究的问题。

关键词:机器学习;随机优化;个体收敛性;有偏梯度估计;非凸问题

中图分类号:TP391　　**文献标志码**:A

Individual Convergence of Stochastic Optimization Methods in Machine Learning

Tao Qing[1], Ma Po[1], Zhang Menghan[1], Tao Wei[2]

(1. 11st Department, Army Officer Academy of PLA, Hefei, 230031, China; 2. College of Command System, The PLA University of Science and Technology, Nanjing, 210007, China)

Abstract: The stochastic optimization algorithm is one of the state-of-the-art methods for solving large-scale machine learning problems, where the focus is on whether or not the optimal convergence rate is derived and the learning structure is ensured. So far, various kinds of stochastic optimization algorithms have been presented for solving the regularized loss problems. However, most of them only discuss the convergence in terms of the averaged output, 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 is extensively explored as an open problem. On the other hand, the commonly-used assumption about unbiased gradient in stochastic optimization often does not hold in practice. In such cases, an astonishing fact is that the bias in the convergence bound of accelerated algorithms will accumulate with the iteration, and this makes the accelerated algorithms inapplicable. In this paper, an overview of the state-of-the-art and existing problems about the stochastic first-order gradient methods is given, which includes the individual convergence rate, biased gradient and nonconvex problems. Based on it, some interesting problems for future research are indicated.

Key words: machine learning; stochastic optimization; individual convergence; biased gradient estimation; non-convex problems

---

基金项目:国家自然科学基金(61673394,61273296)资助项目。
收稿日期:2016-12-05;**修订日期**:2017-01-10

引 言

通俗地说，机器学习的主要目的是让计算机系统具有类似于人的学习能力。如何设计基于训练数据的学习算法从有限样本集合得到分布意义下最优的正确率，是统计机器学习研究的主要内容[1]。基于有限样本学习问题泛化界的统计分析，人们在机器学习算法设计中广泛采用正则化损失函数的优化准则[2]为

$$\min r(\boldsymbol{w}) + l(\boldsymbol{w}) \tag{1}$$

式中：$r(\boldsymbol{w})$为正则化项，常用的正则化项有 L1-范数和 L2-范数。$l(\boldsymbol{w}) = \sum_{i=1}^{m} l_i(\boldsymbol{w})$，$l_i(\boldsymbol{w})$为对应样本$(x_i, y_i)$的损失函数，$m$ 为训练样本的数目。正则化 $r(\boldsymbol{w})$项主要用于增强算法的适定性以及对特定数据的适应性，而损失函数项 $l_i(\boldsymbol{w})$主要描述学习的精度。

随着数据规模的急剧增加和各种应用驱动学习范式的不断涌现，如何求解大规模不同学习范式导致的多样性正则化损失函数优化问题已经成为机器学习领域亟需解决的关键性科学问题。实际上，自从支持向量机(Support vector machine，SVM)出现后，学习问题的优化算法一直是机器学习领域众多研究者普遍关注的一项议题。计算机领域内的一些高水平国际会议经常出现以优化算法为主导的 tutorial，机器学习顶级会议 ICML(International conference on machine leaning)，NIPS(Neural information processing systems)和 COLT (Conference on learning theory)也曾多次举办 workshop 进行专门讨论，机器学习顶级刊物 JMLR(Journal of machine learning research)更是为此设立了 special topic 和出版了 special issue。一些在数学规划领域有重要影响的团队也积极投身于大规模机器学习优化问题研究的热潮中，数学规划领域的权威期刊 SIOPT(SIAM journal on optimization)等频频出现求解机器学习优化问题的论文。目前，依靠机器学习自身特点驱动而迅速发展起来的随机优化算法成为解决大规模问题的有效手段。如何在这些优化算法中保持学习问题的正则化项结构以及获取最优收敛速率是研究中的核心问题。在正则化机器学习问题及其优化算法方面，已经发表了很多综述性论文，如文献[3～8]。特别，在文献[7]中以损失函数和优化求解为主线，对统计机器学习的正则化损失函数框架进行了综述和分析，具体讨论了几何 margin 和损失函数两种理论分析方法，重点讲述了损失函数的渐近最优性，阐述了数学优化方法和机器学习优化算法之间的区别和联系，强调了任何关于损失函数和优化求解方法的进展都将会促进机器学习的发展。文献[8]首先指出大规模机器学习问题的训练样本集合往往具有冗余和稀疏的特点，机器学习优化问题中的正则化项和损失函数也蕴含着特殊的结构含义，直接使用整个目标函数梯度的批处理黑箱方法不仅难以处理大规模问题，也无法满足机器学习对结构的要求，然后针对 L1 正则化问题，介绍了依据机器学习自身特点驱动而迅速发展起来的坐标优化、在线和随机优化方法的一些研究进展。目前，很多经典的一阶优化方法经过适当的改造，不仅可以保证正则化项的结构，同时还具有不依赖正则化项光滑性的最优收敛速率，但由于最终解大多都采取了加权平均的输出方式，仍然存在着一些弊端。特别是对 L1 正则化问题，尽管每一步迭代产生的解具有稀疏性，但平均求和的最终输出方式却破坏了这种求解算法最初极力维护的稀疏性。为了更好地保持正则化项的结构，应该研究个体收敛性问题，但即使是对数学规划中讨论的标准凸优化问题，很多经典文献和书籍也都缺乏个体收敛速率的论述[9]。一般地说，时空复杂性是衡量一个算法优劣的主要标准。结合机器学习问题的实际含义，人们评判学习问题(1)优化算法的依据主要是收敛速率的界、正则化项结构的保持以及存储复杂性。因此，本文综述也在首先介绍随机结构优化算法的基础上，按照个体最优速率这一评判标准展开，这也是本文与文献[3～8]的不同之处。另外，当标准随机优化算法一些普遍使用的如样本集合独立同分布假设不成立时，本文还介绍了一些重要的拓广问题。

1　随机优化算法

目前,机器学习领域的研究者们主要借助数学规划特别是凸优化领域的一阶梯度方法,考虑无约束的优化问题

$$\min f(\boldsymbol{w}) \tag{2}$$

式中:$f(\boldsymbol{w})$为定义在 $\mathbf{R}^n$ 上可微的凸函数,记 $\boldsymbol{w}_*$ 是问题(2)的一个最优解。对于问题(2),批处理形式经典的梯度下降方法的迭代步骤为

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-a_t\nabla f(\boldsymbol{w}_t) \tag{3}$$

式中:a_t 为迭代步长,$\nabla f(\boldsymbol{w}_t)$为 $f(\boldsymbol{w})$在 $\boldsymbol{w}_t$ 处的梯度。梯度下降方法主要在每一点处目标函数梯度的相反方向会使目标函数下降最快这一事实。在梯度下降方法的基础上,针对约束或者非光滑问题,优化方法出现了很多变形,其中包括投影次梯度算法[9]、镜像下降[10]、对偶平均[11]和交替方向乘子方法[12]等。但对于正则化加损失函数这种具有明确机器学习含义的优化问题(1)来说,仍然属于黑箱方法。对于求解特定领域的问题,优化理论方面著名学者 Nesterov 曾经指出"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]。因此,为了利用这些经典算法有效处理大规模具有实际含义的机器学习问题,还需要在使用形式上进行必要的改变。

首先,这些批处理形式的优化方法由于每次迭代都要涉及到损失函数梯度的计算,从而不可避免地需要遍历训练样本集合,该操作方式显然无法适用于大规模机器学习问题的求解。对于优化问题(1),梯度下降方法随机形式可表示为

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-a_t\nabla l_i(\boldsymbol{w}_t) \tag{4}$$

式中:i 为从样本集合中第 $t+1$ 次随机抽取样本的序号。广义上来说,随机优化算法的每步迭代仅需要知道目标函数梯度的无偏估计,特别对于有限样本的机器学习问题来说,由于假设样本是独立同分布的,单个样本对应目标函数的梯度就是整个训练集上目标函数梯度的无偏估计,从而随机优化方法只需要计算部分甚至单个样本对应目标函数的梯度,这就克服了批处理算法每次迭代都需要遍历训练样本集合的固有缺陷。另一方面,由于大规模学习问题训练样本往往存在着冗余现象,实际应用中往往只需运行随机优化算法少许迭代步骤后,学习精度就已经呈现出稳定的趋势[13]。2007 年,Shalev-Shwartz 使用随机投影次梯度对大规模 SVM 进行求解(称为 Pegasos)[14],取得了轰动一时的实际效果。即使是在 10 年后的今天,很多新算法的比较对象中仍然还会有 Pegasos 的身影出现。

其次,文献[15]指出,当将 L1 正则化项和损失函数整体作为目标函数使用一阶梯度随机优化方法时,却无法获得 L1 正则化项应该带来的稀疏性,这表明黑箱优化方法难以保证优化问题中正则化项的结构。为了解决这一问题,Xiao 等在 2009 年提出了一种对保持正则化项结构具有重要影响的算法称之为 RDA(Regularized dual average),成功地将对偶平均优化算法推广至正则化情形[16]。随后,Duchi 等将镜像下降算法推广至正则化情形,称之为 COMID(Composite objective mirror descent)[17]。对于问题(2),批处理形式镜像下降方法的迭代步骤为

$$\boldsymbol{w}_{t+1}=\arg\min_{w\in Q}\{a_t\langle\nabla f(\boldsymbol{w}_t),\boldsymbol{w}\rangle+\boldsymbol{B}(\boldsymbol{w},\boldsymbol{w}_t)\} \tag{5}$$

式中:函数 $\boldsymbol{B}(\boldsymbol{w},\boldsymbol{w}_t)$为 Bregman divergence,特别可取 $\boldsymbol{B}(\boldsymbol{w},\boldsymbol{w}_t)=\frac{1}{2}\|\boldsymbol{w}-\boldsymbol{w}_t\|^2$。对于问题(1),COMID 算法的迭代步骤为

$$\boldsymbol{w}_{t+1}=\arg\min_{w\in Q}\{r(\boldsymbol{w})+a_t\langle\nabla l_i(\boldsymbol{w}_t),\boldsymbol{w}\rangle+\boldsymbol{B}(\boldsymbol{w},\boldsymbol{w}_t)\} \tag{6}$$

当 $r(\boldsymbol{w})$为 L1 范数时,不难发觉 COMID 在优化过程中将正则化项和损失函数分别看待,只是对损失函数采用相应的一阶梯度方法,从而很好地保证了 L1 正则化问题每一步迭代解的稀疏性。这样的研究思路目前已被广泛采用,有时人们也将使用了这种技巧的优化方法称为算法的 Proximal 形式[18-19]或

者更加明确地称为结构优化方法[16]。除了能够保证正则化项的结构之外,结构优化方法的重要理论意义还在于其收敛速率仅与损失函数的光滑性有关,即尽管目标函数非光滑,但只要损失函数满足光滑性条件,结构优化方法仍然能够获得目标函数光滑时才能达到的最优收敛速率,这表明结构优化方法的最优收敛速率界比黑箱方法有数量级的提升。因此,理论分析和实际应用均表明经典优化方法的随机结构化形式在处理大规模学习问题方面具有独特的优势。值得指出的是,在近期优化算法的专题讨论中,经典优化方法的随机结构化形式占了相当大的比重,一些新的研究动向与进展甚至对经典数学规划理论本身的发展产生了积极的影响。

2 最优收敛速率

众所周知,对于不同凸性和光滑性的优化问题,确定型批处理形式优化算法的最优收敛速率已经有了很多定论[20]。显然,评判在机器学习中起重要作用的随机结构优化算法的收敛性是理论分析首要面对的问题。对非光滑损失函数问题,镜像下降方法和对偶平均方法均已推广至随机结构化情形,RDA 和 COMID 对于无强凸性假设的一般凸损失函数情形均能达到最优的收敛速率[16-17]。总体来说,非光滑一般凸问题的研究较为平淡,在获取平均方式输出的最优收敛速率界方面,很多算法都没有碰到特别棘手的困难。相比于非光滑情形,光滑损失函数的优化研究比较丰富。这主要是因为当目标函数光滑时,存在着一种加速技巧,可以将梯度方法的收敛速率提升一个数量级。1983 年,Nesterov 首先提出了这种里程碑式的加速方法[21],获得了目标函数光滑时批处理优化算法的最优收敛速率。随后,这种技巧出现了多种形式的变形[18,22]。因此,大量的随机优化研究都围绕光滑损失函数优化问题的加速上,Nesterov 的加速方法及其变形也已成功地推广至随机结构化情形[16,23]。这些较早期的随机优化算法收敛性分析几乎都是先建立相应在线算法的 regret 界,随后通过在线与随机算法之间的切换技巧[24],得到随机算法对所有迭代进行平均作为输出方式的收敛速率。但是,这种通过在线与随机切换而获得收敛速率的方法在处理强凸优化时却遭遇了一些困境。这是因为当目标函数强凸时,即使是采用对所有迭代进行平均的输出方式,目前标准的 SGD(Stochastic gradient descent) 也未能证明能获得最优的收敛速率界。由此一些学者甚至对 SGD 在经典优化理论中的统治地位产生了质疑,强凸问题优化算法最优收敛速率问题也引起了研究者的普遍关注。相比于一般凸问题,强凸目标函数的收敛速率研究可谓是精彩纷呈。笼统地说,存在着两种主要的思路:(1)对算法的迭代过程进行修改,如 Hazan 等在 SGD 中嵌入适当数目的内循环,提出一种 Epoch-GD 算法[24],获得了平均输出方式的最优收敛速率。(2)仅对输出方式进行修改,如 Rakhlin 等以后半部分迭代解平均的 α-suffix 技巧来代替全部平均的输出方式[25],Lacoste-Julien 等提出了一种加权平均的输出方式[26],这些输出方式均使得 SGD 获得了最优的收敛速率,其中加权平均克服了 α-suffix 技巧输出不能 on-the-fly 计算的问题。

上述讨论的优化方法不仅获得了最优收敛速率,而且计算复杂性和收敛速率的界还与样本数目无关,因此理论上不需要存储样本集合,可以解决任意规模的问题。此时,也可以认为在训练样本独立同分布的假设下,这些算法所优化目标函数中含有基于损失函数的期望风险[2,19]。但是在这种情形下,即使是黑箱方法,当目标函数强凸且光滑时,标准的随机梯度优化方法只能获得次线性的收敛速率界,与全梯度方法的线性收敛速率未能很好匹配,这种差异说明仅仅依靠目标函数梯度的无偏估计就期望得到全梯度方法的效果不现实[19,27],还需要对优化问题进行一定的限制。

在实际应用中,通常给定样本集合,此时的机器学习优化问题表现为正则化经验风险的形式。通过充分利用经验风险这种"有限和"形式的结构,优化算法往往可以得到更好的收敛速率界。自然,处理这种"有限和"形式目标函数的优化算法也可以采取每一步迭代仅对其中一个损失函数进行操作的方式。为了区别起见,将这样的优化算法称为增量算法[9,28],如果以随机方式选取操作的损失函数,此时的算法称为随机增量优化算法。从形式上看,增量算法的迭代方式似乎与 SGD 完全相同,但关键的不同是,

增量算法可以存储部分梯度或迭代信息,这正是增量算法比标准随机方法具有更好收敛速率界的主要原因。

关于"有限和"形式增量算法的研究成果层出不穷,其中代表性的论文有 SAG[27,29],SAGA[30],Finito[28]和 MISO[31]等。不论是处理黑箱还是 Proximal 形式的问题,当目标函数强凸且光滑(损失函数)时,这些方法均可获得和批处理全梯度算法相同阶的收敛速率界。更进一步,为了避免增量算法在处理大规模问题时的存储瓶颈,Johnson 等将减小方差策略嵌入到 SGD 中,称为 SVRG[32]。SVRG 的主要思路是使用全梯度对 SGD 迭代中的梯度进行修正,从而达到了减少方差的目的,同时也付出了不得不周期性地遍历样本集合的代价。但对于光滑强凸优化问题,SVRG 在和标准 SGD 具有几乎相同存储需求的条件下,实现了与 SAG 一样的线性收敛速率。随后,SVRG 被成功拓广至 Proximal 情形,对目标函数强凸但损失函数光滑的问题也获得了线性的收敛速率[19]。另外,一些研究者还对强凸与只有光滑条件的增量算法作了统一的处理[33],还有一些研究工作使用 Nesterov 技巧对增量算法也进行了加速,得到了最优收敛速率界[34-35]。

3　个体收敛速率

使用迭代过程中的个体直接作为输出在稀疏学习问题中具有更明确的实际意义,能够比平均方式的解获得更高的稀疏效果。这说明为了更好地保持正则化项的结构,应该研究个体收敛性问题。但即使是对数学规划中专门讨论的投影次梯度批处理优化方法,如果没有目标函数光滑性假设保证下的单调性[18,22],很多经典文献和书籍也都缺乏个体收敛速率的论述[9]。因此,个体收敛性在数学规划理论研究中也有其重要的地位。下面对个体收敛速率的主要进展和存在的问题进行简单的梳理。

对于标准随机优化算法 SGD,只有当目标函数不仅强凸而且光滑时,才能相对比较容易地获得个体收敛速率界[25]。对于单纯的强凸问题,如果不改变算法本身,也必须对平均的输出方式进行修改,才能获得最优收敛速率[25-26],但研究者对这些结果似乎不满意。实际上,人们最为期待的莫过于 SGD 对于强凸问题能否达到最优个体收敛速率,这个问题看似简单,截至目前却始终没有答案。为了捍卫 SGD 的经典与尊严,Shamir 在 2012 年的机器学习顶级会议 COLT 上把强凸目标函数下 SGD 的个体最优收敛速率作为 open 问题提出[36]。2013 年,Shamir 等提出一种由平均输出方式收敛速率得到个体收敛速率的一般技巧[37],尽管得到了众盼所归的 SGD 个体收敛速率,但获得的收敛速率界与平均输出方式的收敛速率界相差一个对数因子,显然未能达到最优收敛速率界。在随机优化算法的个体收敛速率研究方面,Chen 等提出的 Optimal RDA 获得了比较全面而又非常理想的结果[38]。其主要的思路是对对偶平均方法进行改进,在每一步迭代中增加一个不同形式的子优化问题求解,对研究者通常独立讨论的一般凸、强凸或光滑等类型的问题,均获得了个体最优收敛速率。2015 年,Nesterov 等在对偶平均方法的迭代中巧妙地嵌入了一种线性插值操作,证明了该方法在一般凸情形下具有最优的个体收敛速率,并且这种个体收敛呈现出与平均方式收敛同样的稳定性[39]。从理论角度来说,这种改动与标准的对偶平均方法区别极小,是对偶平均方法一种很好的扩展,也是对一阶梯度方法个体最优收敛速率比较接近大家期待的一种回答。

仔细分析可以发现,文献[38]中的 Optimal RDA 和文献[39]中的线性插值操作技巧获得个体收敛速率的原理不同,以至于获得的学习结果也不相同。一般地说,在处理 L1 正则化问题时,Optimal RDA 的个体解具有很好的稀疏性,但却不具有收敛的稳定性,这也是子优化问题的解直接作为最终输出的通用弊病;而线性插值技巧嵌入在迭代过程的梯度运算后,从而使最终的个体解具有很好的收敛稳定性,但却不具有稀疏性,这也是将插值累积作为最终输出的通用弊病。另外,这两种个体收敛速率分析的思路目前仅适用于步长策略灵活的对偶平均方法。值得指出,文献[38]对对偶平均算法的改动很大,这实际上与标准 SGD 个体收敛速率 open 问题的本意已有所偏离,而线性插值操作技巧[39]对于强凸或光滑

等目标函数情形能否得到个体最优收敛速率也未讨论。最近,文献[40]提出了一种嵌入线性插值操作的投影次梯度方法,证明了其在一般凸情形下具有个体最优收敛速率。更进一步,将所获结论推广至随机方法情形。

4 随机优化算法的拓广

目前大多数的机器学习随机优化方法研究都假设随机抽取的部分甚至单个样本点对应的目标函数梯度是整个目标函数梯度的无偏估计。在这种假设条件下,特别是对光滑损失函数的优化问题,如前所述,人们得到了众多形式的加速算法,这些算法具有最优的收敛速率且收敛速率的界与样本数目无关。

但在实际问题中,梯度无偏估计的假设往往是不成立的。(1)当在给定训练集合的条件下企图优化基于损失函数期望形式的目标函数时,根本无法知晓样本集合是否满足独立同分布条件,此时当然认为梯度估计出现有偏更为合理。(2)L1 正则化问题是一种最为简单典型的正则化损失函数优化问题,其简单之处在于在每一步迭代过程中涉及的优化子问题可以解析求解,从而整个随机算法具有非常理想的计算代价,但对于 Total-variation, Nuclear-norm 等类型的正则化项和 Fused-Lasso 等问题,子问题只能近似求解,这种求解方式也可以视为是一种梯度估计有偏情形下的精确求解[41]。(3)在光滑损失函数的优化算法中,需要知道损失函数的 Lipschitz 常数,这个决定步长的参数对算法性能会产生严重的影响,Lipschitz 常数估计中的误差问题也可以归结为一种广义的梯度无偏估计问题[42]。优化领域著名学者 Nesterov 领导的研究小组对梯度有偏情形进行了深入的研究,他们定义一般形式的梯度无偏估计问题,目前很多文献中所关注的具体情形都是其特例[42]。一个令人惊讶的事实是所有研究者对所涉及的具体梯度估计有偏问题都得到了一致相同的结论,即不论是批处理还是随机优化算法,非加速算法收敛速率界保持梯度偏差为一常量,而加速算法平均输出形式收敛速率的界却会出现随着迭代的增加而累加定性误差的现象,这使得加速算法无任何理论保证,处于十分尴尬的境地[41,43-44]。对于非加速算法,在假定了广义梯度偏差以一定速率衰减的条件下,可以获得通常意义下极限为 0 的收敛速率界,且这个结论可以用来求解一种矩阵分解问题,其中的偏差由原问题和对偶问题之间的间隙控制[41]。

对于正则化损失函数形式的优化问题,有时人们对正则化项或损失函数项强加了更多的学习含义。例如,为了获得更稀疏的支持向量,人们对 hinge 损失采取了强行截断的手段[45-46];同样地,为了获得更好的特征稀疏度,人们对 L1 正则化项也采取了截断的手法[47]。与标准的凸优化模型相比,这些截断模型也具有很好的统计含义,并体现了正则化项和损失函数对学习问题的原本目标更精确的逼近。但不可回避的是,这些模型都带来一个难以处理的问题,即如何求解这些对学习问题有额外需求而导致的非凸优化问题。非凸优化问题一直是数学规划领域的难解问题,只是在一些特定的假设下,才能得到一些局部收敛的算法。MM(Majorization-minimization)方法可以视为处理非凸优化问题的一般框架,在迭代过程中,MM 算法通过优化目标函数局部凸上界的形式避开了目标函数的非凸性[48]。著名算法 CCCP(Concave convex procedure)[49]或 DC (Difference convex)规划[50]实际上是求解两个凸函数之差形式非凸问题的特例,并且人们已经使用 CCCP 和 DC 规划处理截断批处理形式的优化问题,均获得了比求解凸优化问题更令人满意的学习结果[45-47],尤其是关于非线性截断 SVM 的求解[45],由于支持向量数目的减少,大大缩短了求解优化问题的时间,这篇论文也获得了 ICML2006 的最佳论文奖。自然,人们会想到能否在非凸批处理算法的基础上进一步拓广随机优化方法求解大规模非凸优化问题。2015 年,针对"有限和"形式的非凸优化问题,受 SAG 方法[27]的启发,Mairal 给出了针对光滑目标函数增量形式的 MM 算法[31]。梯度有偏情形的大规模优化问题和非凸问题的研究同属于凸优化问题随机优化算法应用范围的一种拓广。目前,也只是在比较严格的假设下,才获得了与理想情形类似的收敛速率界,关于个体收敛速率界的结果还未见报道,非"有限和"形式目标函数的探讨更是极少。

5　结束语

综上所述,随机优化算法是求解大规模机器学习问题的主流优化方法之一,具有坚实的理论基础和良好的发展前景,但在大家普遍关注的问题中仍然存在着很多令人无法回避而又亟待需要克服的缺陷。根据上述对随机优化研究进展的梳理,以下几个方面值得研究:

(1)对偶平均方法经过一些改进,无论是对一般凸、强凸还是光滑的损失函数,已经具有了最优的个体收敛速率,但在收敛稳定性或保持正则化项结构方面仍然不能令人完全满意。因此,能否对目前主流的一阶随机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,这是经典梯度方法在机器学习优化问题应用中必须解决的问题,也是对 SGD 个体收敛速率 open 问题比较接近的回答。

(2)实际问题中往往会出现目标函数梯度估计有偏情形,假设梯度估计有偏是从 L1 正则化损失函数优化问题过渡到一般正则优化问题的一种有效途径,也是处理参数不确定性的一种手段。遗憾的是,无偏情形下一些具有最优收敛速率随机算法的收敛速率界却会随迭代次数累计增加偏差项。因此,一些应用中必须研究的问题,也是个体收敛性研究中值得研究的问题。

(3)由于实际应用的驱动,机器学习领域也出现了一些具有特定含义的非凸正则化问题,目前蓬勃发展并取得空前成功的深度神经网络也强烈地依赖于非凸优化问题的求解[32]。非凸问题研究虽然有一些进展,但还存在相当多的问题没有涉及。因此,研究非凸问题的个体最优收敛速率界与稳定性问题具有理论意义和很好的应用前景。

总之,随机优化方法的个体最优收敛速率、收敛稳定性及其拓广的相关研究具有重要的理论意义和实用价值,在当前机器学习优化方法发展中具有一定的挑战性问题,这项研究对算法的进一步分布式实现也具有指导意义。

参考文献:

[1] Vapnik V N. Statistical learning theory[M]. New York: Wiley-Interscience, 1998.

[2] Zhang T. Statistical behavior and consistency of classification methods based on convex risk minimization[J]. Annals of Statistic, 2004,32:56-85.

[3] Duchi J. Introductory lectures on stochastic convex optimization [EB/OL]. Park City Mathematics Institute, Graduate Summer School Lectures, http://web.stanford.edu/~jduchi/, 2016.

[4] Bottou L, Curtis F, Nocedal J. Optimization methods for large-scale machine learning[EB/OL]. https://arxiv.org/abs/1606.04838, 2016-06-15[2016-06-15].

[5] 吴启晖,邱俊飞,丁国如.面向频谱大数据处理的机器学习方法[J].数据采集与处理,2015,30(4):703-731.
Wu Qihui, Qiu Junfei, Ding Guoru. Maching learning methods for big spectrum data processing[J]. Journal of Data Acquistion and Processing,2015,30(4):703-731.

[6] 潘志松,唐斯琪,邱俊洋,等.在线学习算法综述[J].数据采集与处理,2016,31(6):1067-1082.
Pan Zhisong,Tang Siqi,Qiu Junyang,et al. Survey on online learning algorithms[J]. Journal of Data Acquistion and Processing,2016,31(6):1067-1082.

[7] 孙正雅,陶卿.统计机器学习综述:损失函数与优化求解[J].中国计算机学会通讯,2009,5(8):7-14.
Sun Zhengya,Tao Qing. The Loss function and optimizer in statistical machine learning[J]. Communications of the Chinese Computer Federation,2009,5(8):7-14.

[8] 陶卿,高乾坤,姜纪远,等.稀疏学习优化问题的求解综述[J].软件学报,2013,24(11):2498-2507.
Tao Qing, Gao Qiankun, Jiang Jiyuan, et al. Survey of solving the optimization problems for sparse learning[J]. Journal of Software,2013,24(11):2498-2507.

[9] Bertsekas D P, Nedić A, Ozdaglar A E. Convex analysis and optimization[M]. Belmont: Athena Scientific, 2003.

[10] Beck A, Teboulle M. Mirror descent and nonlinear projected sub-gradient methods for convex optimization[J]. Oper Res

Lett, 2003,31:167-175.

[11] Nesterov Y. Primal-dual subgradient methods for convex problems[J]. Math Program, 2009,120:261-283.

[12] Boyd S, Parikh N, Chu E, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers[J]. Foundations and Trends in Machine Learning, 2011, 3(1):1-122.

[13] Zhang T. Solving large scale linear prediction problems using stochastic gradient descent algorithms[C]//Proceedings of the International Conference on Machine Leaning. New York: ACM, 2004: 919-926.

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

[15] Langford J. Sparse online learning via truncated gradient[J]. Journal of Machine Learning Research, 2008, 10(2):777-801.

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

[17] Duchi J, Shalev-Shwartz S, Singer Y, et al. Composite objective mirror descent[C]// Proceedings of the Conference on Learning Theory. Haifa, Israel:[s. n.],2010:14-26.

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

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

[20] Nemirovsky A S, Yudin D B. Problem complexity and method efficiency in optimization[M]. New York: Wiley, 1983.

[21] 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.

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

[23] Hu C, Kwok J T, Pan W. Accelerated gradient methods for stochastic optimization and online learning[C]//Advances in Neural Information Processing Systems. Vancouver, British Columbia, Canada:[s. n.] ,2009:781-789.

[24] Hazan E, Kale S. Beyond the regret minimization barrier: An optimal algorithm for stochastic strongly-convex optimization [J]. Journal of Machine Learning Research, 2011,15(1):2489-2512.

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

[26] Lacoste-julien S, Schmidt M, Bach F. A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method[EB/ OL] . https://arxiv. org/abs/1212. 2002, 2012-12-20[2012-12-20].

[27] Schmidt M, Roux N L, Bach F. Minimizing finite sums with the stochastic average gradient[J]. Mathematical Programming, 2013,26(5):405-411.

[28] Defazio A J, Caetano T S, Domke J. Finito: A faster, permutable incremental gradient method for big data problems[C]// Proceedings of the International Conference on Machine Leaning. Beijing:[s. n.], 2014:1125-1133.

[29] Roux N L, Schmidt M, Bach F R. A stochastic gradient method with an exponential convergence _rate for finite training sets [C]//Advances in Neural Information Processing Systems. Lake Tahoe:[s. n.], 2012: 2663-2671.

[30] Defazio A, Bach F, Lacoste-julien S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives[C]//Advances in Neural Information Processing Systems. Canada:[s. n.] ,2014:1646-1654.

[31] Mairal J. Incremental majorization-minimization optimization with application to large-scale machine learning[J]. SIAM Journal on Optimization, 2015,25(2):829-855.

[32] Johnson R, Zhang T. Accelerating stochastic gradient descent using predictive variance reduction[C]// Advances in Neural Information Processing Systems. Lake Tahoe: [s. n.], 2013:315-323.

[33] Allen Zhu Z , Yuan Y. Univer: A universal variance reduction framework for proximal stochastic gradient method[EB/OL]. https://arxiv. org/abs/1506. 01972, 2015-07-05[2015-07-05].

[34] Atsushi N. Stochastic proximal gradient descent with acceleration techniques[C]//Advances in Neural Information Processing System. Montreal, Canada:[s. n.], 2014:1574-1582.

[35] Lin H, Mairal J, Harchaoui Z. A universal catalyst for first-order optimization[C]//Advances in Neural Information Process-

ing System. Montreal,Canada:[s. n.], 2015:1604-1616.
[36] Shamir O. Open problem: Is averaging needed for strongly convex stochastic gradient descent[C]//Proceedings of the Conference on Learning Theory. USA:MIT Press,2012:471-475.
[37] Shamir O, Zhang T. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes[C]//Proceedings of the International Conference on Machine Learning. Atlanta,USA:[s. n.],2013:71-79.
[38] Chen X, Lin Q, Pena J. Optimal regularized dual averaging methods for stochastic optimization[C]//Advances in Neural Information Processing Systems. Lake Tahoe: [s. n.], 2012, 404-412.
[39] Nesterov Y, Shikhman V. Quasi-monotone subgradient methods for nonsmooth convex minimization[J]. J of Opti Theory and Appl, 2015, 165(3): 917-940.
[40] 陶蔚,潘志松,储德军,等.线性插值投影次梯度方法的最优个体收敛速率[J].计算机研究与发展,2017,54(3):1-8.
Tao Wei, Pan Zhisong, Chu Dejun, et al. The optimal individual convergence rate for the projected subgradient methods with linear interpolation operation[J]. Journal of Computer Research and Development, 2017,54(3):1-8.
[41] Schmidt M, Le Roux N, Bach F. Convergence rates of inexact proximal-gradient methods for convex optimization[J]. Advances in Neural Information Processing Systems, 2011, 24:1458-1466.
[42] Devolder O, Glineur F, Nesterov Y. First-order methods of smooth convex optimization with inexact oracle[J]. Mathematical Programming, 2014, 146(1/2):37-75.
[43] Devolder O. Stochastic first order methods in smooth convex optimization[EB/OL]. CORE Discussion Papers, http://www.uclouvain.be/en-357992.html, 2011.
[44] Honorio J. Convergence rates of biased stochastic optimization for learning sparse ising models[C] //Proceedings of the International Conference on Machine Learning. Edinburgh, Scotland:[s. n.], 2012:257-264.
[45] Collobert R, Sinz F, Weston J, et al. Trading convexity for scalability[C]// Proceedings of the International Conference on Machine Learning. USA: MIT Press, 2006:201-208.
[46] Wu Y, Lin Y. Robust truncated hinge loss support vector machines[J]. Journal of the American Statistical Association, 2007,102(479): 974-983.
[47] Zhang T. Mutil-stage convex relaxation for learning with sparse regularization[C]// Advances in Neural Information Processing System. Vancouver, British Columbia, Canada:[s. n.],2008:1929-1936.
[48] Lange K, Hunter D R, Yang I. Optimization transfer using surrogate objective functions[J]. J Comput Graph Statist, 2000 (9):1-20.
[49] Yuille A L, Rangarajan A. The concave-convex procedure[J]. Neural Computation, 2003,15(4):915-936.
[50] Horst R, Thoai N V. DC programming: Overview[J]. J Optim Theory Appl, 1999,103:1-43.

作者简介:

陶卿(1965-),男,教授,博士生导师,研究方向:机器学习、模式识别和应用数学,E-mail:qing.tao@ia.ac.cn,taoqing@gmail.com。

马坡(1993-),男,硕士研究生,研究方向:凸优化及其在机器学习中的应用。

张梦晗(1994-),男,硕士研究生,研究方向:模式识别、人工智能。

陶蔚(1991-),男,硕士研究生,研究方向:机器学习、网络安全和模式识别。