DOI: 10.11992/tis.202006046

非光滑凸情形 Adam 型算法的最优个体收敛速率

黄鉴之[1], 丁成诚[1], 陶蔚[2], 陶卿[1]

（1. 中国人民解放军陆军炮兵防空兵学院 信息工程系, 安徽 合肥 230031; 2. 中国人民解放军陆军工程大学 指挥控制工程学院, 江苏 南京 210007）

摘　要: Adam 是目前深度神经网络训练中广泛采用的一种优化算法框架, 同时使用了自适应步长和动量技巧, 克服了 SGD 的一些固有缺陷。但即使对于凸优化问题, 目前 Adam 也只是在线学习框架下给出了和梯度下降法一样的 regret 界, 动量的加速特性并没有得到体现。这里针对非光滑凸优化问题, 通过巧妙选取动量和步长参数, 证明了 Adam 的改进型具有最优的个体收敛速率, 从而说明了 Adam 同时具有自适应和加速的优点。通过求解 l_1 范数约束下的 hinge 损失问题, 实验验证了理论分析的正确性和在算法保持稀疏性方面的良好性能。

关键词: 机器学习; AdaGrad 算法; RMSProp 算法; 动量方法; Adam 算法; AMSGrad 算法; 个体收敛速率; 稀疏性

中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1673−4785(2020)06−1140−07

中文引用格式: 黄鉴之, 丁成诚, 陶蔚, 等. 非光滑凸情形 Adam 型算法的最优个体收敛速率 [J]. 智能系统学报, 2020, 15(6): 1140–1146.

英文引用格式: HUANG Jianzhi, DING Chengcheng, TAO Wei, et al. Optimal individual convergence rate of Adam-type algorithms in nonsmooth convex optimization[J]. CAAI transactions on intelligent systems, 2020, 15(6): 1140–1146.

Optimal individual convergence rate of Adam-type algorithms in nonsmooth convex optimization

HUANG Jianzhi[1], DING Chengcheng[1], TAO Wei[2], TAO Qing[1]

(1. Department of Information Engineering, Army Academy of Artillery and Air Defense of PLA, Hefei 230031, China; 2. Command and Control Engineering, Army Engineering University of PLA, Nanjing 210007, China)

Abstract: Adam is a popular optimization framework for training deep neural networks, which simultaneously employs adaptive step-size and momentum techniques to overcome some inherent disadvantages of SGD. However, even for the convex optimization problem, Adam proves to have the same regret bound as the gradient descent method under online optimization circumstances; moreover, the momentum acceleration property is not revealed. This paper focuses on nonsmooth convex problems. By selecting suitable time-varying step-size and momentum parameters, the improved Adam algorithm exhibits an optimal individual convergence rate, which indicates that Adam has the advantages of both adaptation and acceleration. Experiments conducted on the l_1-norm ball constrained hinge loss function problem verify the correctness of the theoretical analysis and the performance of the proposed algorithms in keeping the sparsity.

Keywords: machine learning; AdaGrad algorithm; RMSProp algorithm; momentum methods; Adam algorithm; AMSGrad algorithm; individual convergence rate; sparsity

Adam 是目前深度学习中广泛采用的一种优化算法[1]。与经典梯度下降不同的是, Adam 同时使用了自适应步长和动量两种技巧。其中自适应步长技巧使算法对超参数不敏感, 动量技巧可以加速算法在处理凸优化问题时的收敛速率, 在处理非凸问题时帮助算法避开鞍点甚至局部极值点。与仅使用单一技巧的方法相比, Adam 在典

收稿日期: 2020−06−28.
基金项目: 国家自然科学基金项目 (61673394; 62076252).
通信作者: 陶卿. E-mail: qing.tao@ia.ac.cn.

型卷积深度学习的实验中具有一定的优势[1]。

在梯度下降方法的基础上, 首先使用基于梯度的自适应步长方法是 AdaGrad, 其主要的思路是通过对过往所有梯度的外积进行累加的方式进行步长自适应的调整[2]。虽然 AdaGrad 在求解非光滑凸问题时具有和投影次梯度方法一样 $O\left(\sqrt{t}\right)$ 的 regret 界, 其中 t 是算法的迭代步数[3]。AdaGrad 算法最初主要用于正则化学习问题求解, 但在深度学习应用中的效果却很不理想。出现这一问题的主要原因是算法对过去梯度的外积单纯做和, 导致累加项变大得过快。为了克服这一缺陷, Hinton 等[4] 采用 EMA (exponential moving average) 的形式修改 AdaGrad 算法累加项的计算方式, 提出了 RMSProp 算法。RMSProp 算法丢弃了相对遥远的历史梯度信息, 保证了若干次迭代后学习能继续进行, 较好地适应了目标函数平滑性的局部变化。在 RMSProp 算法的基础上, 自适应步长算法又有了进一步的发展, 典型的算法有 AdaDelta[5] 等。

动量法是在经典梯度下降法基础上通过添加动量项演变而来的, 广泛用于提高一阶梯度算法的收敛速率。根据动量表示方式的不同以及动量项中计算梯度位置的不同可以分成两类: 一类是 Polyak[6] 于 1964 年提出的 Heavy-ball 型动量法; 另一类是 Nesterov[7] 于 1983 年提出的 NAG (nesterov accelerated gradient) 型动量法。其中, Heavy-ball 直接在当前点计算梯度; 而 NAG 会根据当前动量预判下次迭代可能抵达的位置, 并在预判的位置计算梯度。动量方法的加速性能主要体现在求解凸光滑目标函数时 NAG 取得的突破, 将梯度下降法的收敛速率 $O(1/t)$ 加速至最优的 $O(1/t^2)$[7]。当目标函数光滑且强凸时, 虽然动量方法和梯度下降法都能达到线性收敛, 但动量法由于具有小的收敛因子仍然具有加速性[8]。

对于非光滑凸优化问题, 投影次梯度方法目前获得的最好的个体收敛速率只是 $O\left(\log(t)/\sqrt{t}\right)$[9]。2018 年, 陶蔚等[10-11] 将 NAG 步长策略引入到投影次梯度中, 得到了最优个体收敛速率 $O\left(1/\sqrt{t}\right)$, 同时保证了良好的稀疏性。2019 年, 程禹嘉等[12] 证明了 Heavy-ball 型动量法具有 $O\left(1/\sqrt{t}\right)$ 的最优个体收敛速率。由此可以看出, 动量方法对非光滑凸问题同样具有加速作用, 只不过体现在个体收敛速率上。从 Heavy-ball 型动量法加速个体速率的证明过程来看, 主要是借鉴了 Ghadimi 等[8] 在光滑条件下 Heavy-ball 算法收敛速率的证明。

Adam 的最初形式是在梯度下降的基础上使用了 EMA 策略修正步长和 Heavy-ball 型动量法修正梯度方向[1]。尽管 Adam 在深度学习中有着不错的表现[13-14], 但自从 Kingma 等于 2015 年提出 Adam 后, 其收敛性一直是一个具有挑战性的问题。对于非光滑凸问题, 尽管 Kingma 等声称证明了 Adam 取得了 $O\left(\sqrt{t}\right)$ 的 regret 界, 但 Reddi 等却对 Adam 的收敛性提出了质疑, 他们通过理论和实验证明, 即使是对简单的一维凸函数, Adam 也会出现无法收敛到最优解甚至收敛到最差局部极值点的现象[15]。出现这一问题的原因在于: 采用 EMA 方式处理梯度时, 在迭代后期步长会出现不降反升的现象, 从而会导致目标函数值剧烈波动。为了克服这一问题, Reddi 等提出了 AMSGrad 和 AdamNC 两个修改版本。其中, AMSGrad 在 Adam 自适应矩阵上添加了一个确保步长衰减的操作, 从而使得在线 AMSGrad 在一般凸的情况下能获得 $O\left(\sqrt{t}\right)$ 的 regret 界。2019 年, Wang 等[16] 通过对 Adam 进行适当的变形, 得到了强凸情形下 $O(\log(t))$ 的 regret 界。

从目前 Adam 的各种发展趋势来看, 我们更愿意将 Adam 视为一种利用过去梯度信息同时更新下降方向和步长的一阶梯度算法框架[17]。本文主要研究非光滑凸情形 Adam 型方法 AMSGrad 的个体收敛速率问题。为了避免叙述上的复杂性, 直接简称为 Adam 算法。

本文的主要工作有以下 3 个方面:

1) 证明了 Adam 具有 $O\left(1/\sqrt{t}\right)$ 的最优个体收敛速率。据我们所知, 这一理论结果填补了 Adam 算法在非光滑条件下个体最优收敛性方面研究的缺失, 同时也说明了 Adam 继承了 Heavy-ball 型动量法的优点, 体现了动量技巧的特性, 可以将个体收敛加速至最优。

2) 本文的收敛性分析思路具有一定的一般性, 本文首先借鉴了 Ghadimi 等[8] 在光滑条件下 Heavy-ball 算法收敛速率的证明方法, 得到与梯度下降法相似的迭代公式。为了进一步得到非光滑条件下的最优个体收敛速率, 与文献 [12] 类似, 本文巧妙设计了时变的步长和动量权重参数, 同时使用 Zinkevich 在处理在线优化问题收敛性时使用的技巧, 处理变步长与权重导致的递归问题[3]。

3) 本文选择了典型的 l_1 范数约束下的 hinge 损失函数优化问题, 通过与几种具有最优收敛速率算法的比较, 验证了理论的正确性和算法在保持稀疏性方面的良好性能。

1 典型自适应及动量算法的收敛性

考虑有约束优化问题:

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

式中: $f(\boldsymbol{w})$ 为凸函数; $Q\subseteq\mathbf{R}^n$ 为有界闭凸集。记 $\boldsymbol{w}^*$ 是式 (1) 的一个最优解。

梯度下降法是解决问题式 (1) 的经典方法, 基于梯度反方向总是指向目标函数下降的方向这一事实, 具体迭代步骤为

$$\boldsymbol{w}_{t+1}=\pi_Q(\boldsymbol{w}_t-\alpha_t\boldsymbol{g}_t)=\arg\min_{\boldsymbol{w}\in Q}\|\boldsymbol{w}-(\boldsymbol{w}_t-\alpha_t\boldsymbol{g}_t)\|_2^2 \tag{2}$$

式中: π_Q 为集合 Q 上的投影算子[18]; $\alpha_t=\alpha/\sqrt{t}$, α 为大于 0 的常数; $\boldsymbol{g}_t$ 是函数 $f(\boldsymbol{w})$ 在点 $\boldsymbol{w}_t$ 处的次梯度。

平均收敛速率指的是 $f(\overline{\boldsymbol{w}}_t-\boldsymbol{w}^*)$ 的收敛速率, 其中 $\overline{\boldsymbol{w}}_t=\left(\sum_{j=1}^{t}\boldsymbol{w}_j\right)\Big/t$。与之对应, 个体收敛速率指的是 $f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)$ 的收敛速率。通常来说, 对于非光滑问题, 个体收敛更难获得最优速率。

Agarwal 等[19] 证明非光滑一般凸情形下投影次梯度式 (2) 能获得 $O\left(1/\sqrt{t}\right)$ 的最优平均收敛速率, 即

$$E[f(\overline{\boldsymbol{w}}_t)-f(\boldsymbol{w}^*)]\leqslant O\left(1/\sqrt{t}\right)$$

AdaGrad 的具体迭代步骤为

$$\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\alpha\boldsymbol{V}_t^{-1/2}\boldsymbol{g}_t$$

式中: $\boldsymbol{V}_t$ 为 $d\times d$ 维对角矩阵, 将矩阵第 i 维上的元素表示为 $\boldsymbol{V}_{t,i}$, $\boldsymbol{V}_{t,i}=\sum_{j=1}^{t}\boldsymbol{g}_{j,i}^2$ 为过去所有梯度第 i 个元素的算数平方和, $\boldsymbol{V}_t^{-1/2}$ 矩阵中对角元素 $\boldsymbol{V}_{t,i}^{-1/2}=\left(\sum_{j=1}^{t}\boldsymbol{g}_{j,i}^2\right)^{-1/2}$。显然, 可以将 AdaGrad 算法的步长理解为 $\alpha\boldsymbol{V}_t^{-1/2}$, 因为梯度累积的影响, 算法会自动增加稀疏维度的步长, 减小其他维度的步长。对于凸的目标函数, AdaGrad 能取得 $O\left(\sum_{i=1}^{d}\left\|\boldsymbol{g}_{1:t,i}\right\|_2\right)$ 的 regret 界, 其中 $\boldsymbol{g}_{1:t}$ 为 $\{\boldsymbol{g}_1,\boldsymbol{g}_2,\cdots,\boldsymbol{g}_t\}$ 构成的 $d\times t$ 维矩阵, $\boldsymbol{g}_{1:t,i}$ 表示由矩阵 $\boldsymbol{g}_{1:t}$ 第 i 行组成的向量。由此可以看出, AdaGrad 在最坏情况下能取得 $O\left(\sqrt{t}\right)$ 的 regret 界, 当梯度是稀疏时该 regret 界会变得更紧[2]。

RMSProp 算法采用 EMA 的形式改变 AdaGrad 算法中矩阵项算数平方和的计算方式, 克服了当梯度较稠密时步长衰减过快的问题。RMSProp 算法中矩阵的计算方式可以表示为

$$\boldsymbol{V}_t=\beta\boldsymbol{V}_{t-1}+(1-\beta)\operatorname{diag}\left(\boldsymbol{g}_t\boldsymbol{g}_t^{\mathrm{T}}\right)$$

式中: $\beta\in[0,1)$; $\operatorname{diag}(\cdot)$ 表示只保留矩阵对角元素, 其余元素置 0。通过设置不同的 β, 分配给过去梯度的权重会以指数方式衰减, 起到主要作用的仅限于最近的几个梯度, 实际应用中一般取 β=0.9。

相较于梯度下降法, Heavy-ball 型动量法使用动量作为迭代的方向, EMA 形式 Heavy-ball 型动量法可以表示为

$$\boldsymbol{m}_t=\beta_t\boldsymbol{m}_t+(1-\beta_t)\boldsymbol{g}_t\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\alpha_t\boldsymbol{m}_t \tag{3}$$

式中 $\beta_t\in[0,1)$。通过巧妙设置步长和动量参数, Heavy-ball 型动量法具有 $O\left(1/\sqrt{t}\right)$ 的最优个体收敛速率[12]。

Adam 是在 RMSProp 基础上结合 Heavy-ball 型动量技巧发展而来的。具体迭代步骤为

$$\begin{aligned}&\boldsymbol{m}_t=\beta_{1,t}\boldsymbol{m}_t+(1-\beta_{1,t})\boldsymbol{g}_t\\&\boldsymbol{V}_t=\beta_2\boldsymbol{V}_{t-1}+(1-\beta_2)\operatorname{diag}\left(\boldsymbol{g}_t\boldsymbol{g}_t^{\mathrm{T}}\right)\\&\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\alpha_t\boldsymbol{V}_t^{-1/2}\boldsymbol{m}_t\end{aligned} \tag{4}$$

式中: $\alpha_t=\dfrac{\alpha}{\sqrt{t}\beta_{1,t}}$; $\beta_2\in[0,1)$。实际应用中一般取 $\beta_{1,t}$=0.9, β_2=0.99。

为了解决 Adam 的收敛性问题, AMSGrad 在 Adam 自适应矩阵上添加一个使步长衰减的操作, 具体形式为

$$\begin{aligned}&\hat{\boldsymbol{V}}_t=\max\left\{\hat{\boldsymbol{V}}_{t-1},\boldsymbol{V}_t\right\}\\&\boldsymbol{w}_{t+1}=\boldsymbol{w}_t-\alpha_t\hat{\boldsymbol{V}}_t^{-1/2}\boldsymbol{m}_t\end{aligned} \tag{5}$$

改进后的 Adam 算法在 $\beta_{1,t}=\beta_1/t$ 时能取得 $O\left(\sqrt{t}\right)$ 的 regret 界。

比较 Heavy-ball 型动量法的式 (3)、Adam 的式 (4) 和 AMSGrad 的式 (5) 可以看出: 除了多了一个自适应步长的矩阵 $\boldsymbol{V}_t^{-1/2}$ 外, Adam、AMSGrad 的关键迭代步骤和 Heavy-ball 型动量法式 (3) 并无区别。既然自适应步长策略并不影响算法的收敛速率, 这就启发我们: Adam 型算法应该和 Heavy-ball 型动量法一样具有最优的个体收敛速率 $O\left(1/\sqrt{t}\right)$。

2 个体收敛速率分析

本节给出非光滑条件下 Adam 算法在个体最优收敛速率的证明。

为了进行个体收敛性分析, 首先借鉴 Ghadimi 等[8] 在光滑条件下 Heavy-ball 算法收敛速率的证明, 引入加权动量项 $p_t=t(\boldsymbol{w}_t-\boldsymbol{w}_{t-1})$。通常情况下, Adam 动量项的 $\beta_{1,t}$ 参数在实际应用中一般设定为常数, 但对于非光滑问题, 这种方式无法获得个体收敛速率。因此本文改变 Adam 动量项的

计算方式为

$$\boldsymbol{m}_t=\beta_{1,t}\boldsymbol{m}_t+\beta'_{1,t}\boldsymbol{g}_t$$

通过巧妙地选取$\beta_{1,t}$和$\beta'_{1,t}$时变参数(见定理 1), 可以将 Adam 算法的迭代步骤转化为

$$\boldsymbol{w}_{t+1}+p_{t+1}=\boldsymbol{w}_t+p_t-\alpha_t\hat{\boldsymbol{V}}_t^{-1/2}\boldsymbol{g}_t$$

这个关系式和梯度下降法的关键迭代相似, 也和 Heavy-ball 型动量法的关键迭代十分类似。正是基于这个关系式, 梯度下降法的收敛分析思路可以用于 Heavy-ball 型动量法。受文献 [12] 的启发, 巧妙设计了时变步长和动量项参数, 得到了个体收敛速率的递归公式, 为了得到递归后个体收敛速率的界, 这里同样要使用 Zinkevich 在线优化时使用的迭代技巧。与文献 [12] 不同的是, 本文需要处理自适应步长矩阵, 而不是人为指定的步长, 这里我们又借鉴 Kingma 在证明在线 Adam 的 regret 界时使用的技巧。

引理 1　令 $D_\infty=\max\limits_{\boldsymbol{w},\boldsymbol{u}\in Q}\|\boldsymbol{w}-\boldsymbol{u}\|_\infty$, $G_\infty=\max\limits_t\|\boldsymbol{g}_t\|_\infty$, 设 $\boldsymbol{w}_t$ 由式 (5) 生成, 则

$$\sum_{j=1}^{t}\sqrt{j}\left\{\left\|\boldsymbol{w}_j-\boldsymbol{w}^*\right\|^2_{\hat{V}_j^{1/2}}-\left\|\boldsymbol{w}_{j+1}-\boldsymbol{w}^*\right\|^2_{\hat{V}_j^{1/2}}\right\}+$$

$$\sum_{j=1}^{t}\frac{1}{\sqrt{j}}\left\|\boldsymbol{g}_j\right\|^2_{\hat{V}_j^{-1/2}}\leqslant$$

$$D_\infty^2\sqrt{t}\sum_{i=1}^{d}\hat{\boldsymbol{V}}_{t,i}^{1/2}+\frac{2G_\infty}{\sqrt{1-\beta_2}}\sum_{i=1}^{d}\left\|\boldsymbol{g}_{1:t,i}\right\|_2$$

定理 1　设 $f(\boldsymbol{w})$ 是凸函数, $\alpha_t=\alpha/\sqrt{t}$, 取 $\beta_{1,t}=\dfrac{t\sqrt{t}}{(t+2)\sqrt{t-1}}\hat{\boldsymbol{V}}_t^{1/2}\hat{\boldsymbol{V}}_{t-1}^{-1/2}$, $\beta'_{1,t}=\dfrac{1}{t+2}$, $\boldsymbol{w}_t$ 由式 (5) 生成, 则

$$f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)\leqslant\frac{f(\boldsymbol{w}_0)-f(\boldsymbol{w}^*)}{1+t}+$$

$$\frac{D_\infty^2\sqrt{t}}{2\alpha(1+t)}\sum_{i=1}^{d}\hat{\boldsymbol{V}}_{t,i}^{1/2}+\frac{\alpha G_\infty}{(1+t)\sqrt{1-\beta_2}}\sum_{i=1}^{d}\left\|\boldsymbol{g}_{1:t,i}\right\|_2$$

至此, 我们成功得到了 Adam 算法在非光滑情况下的最优个体收敛速率, 然而, 批处理形式的 Adam 算法每次迭代都使用全部样本计算次梯度 $\boldsymbol{g}_t$, 该操作对大规模机器学习问题显然是不可行的。为了解决该问题, 将 Adam 算法推广至随机形式。

本文仅考虑二分类问题, 假设训练样本集 $S=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}\in R^d\times\{+1,-1\}$, 其中 $\boldsymbol{x}$ 和 y 分别对应样本的特征向量和监督值,同时 $(\boldsymbol{x}_i,y_i)$ 之间满足独立同分布。并且只考虑最简单的非光滑稀疏学习问题“hinge 损失”, 即 $f_i(\boldsymbol{w})=\max(0,1-y_i(\boldsymbol{w},\boldsymbol{x}_i))$ 的目标函数为

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

约束情况下随机形式的 Adam 算法迭代步骤可以表示为

$$\boldsymbol{w}_{t+1}=\pi_Q\left(\boldsymbol{w}_t-\alpha_t\hat{\boldsymbol{V}}_t^{-1/2}\boldsymbol{m}_t\right)\tag{6}$$

相对于批处理形式的次梯度 $\boldsymbol{g}_t$, 随机优化每次只选择一个样本计算随机次梯度 $\nabla f_i(\boldsymbol{w})$。由于“hinge 损失”次梯度的计算方式有很多种, 这里选择文献 [20] 的方式:

$$\nabla f_i(\boldsymbol{w})=\frac{1}{m}\sum_{(\boldsymbol{x}_i,y_i)\in A_t^+}y_i\boldsymbol{x}_i\tag{7}$$

其中 $A_t^+\subseteq S$, $A_t^+=\{(\boldsymbol{x}_i,y_i)\subseteq A_t;y_i\langle\boldsymbol{w},\boldsymbol{x}_i\rangle<1\}$, 实验中选取 $|A_t|=1$。

随机 Adam 算法:

1) 初始化 $\boldsymbol{w}_1\in Q$, 定义步长为 α_t 且 $\{\alpha_t>0\}_{t=1}^T$;
2) for t=1 to T;
3) 等可能性选取 i;
4) 根据式 (7) 计算随机次梯度 $\nabla f_i(\boldsymbol{w}_t)$;
5) 取 $\beta_{1,t}=\dfrac{t\sqrt{t}}{(t+2)\sqrt{t-1}}\hat{\boldsymbol{V}}_t^{1/2}\hat{\boldsymbol{V}}_{t-1}^{-1/2}$, $\beta'_{1,t}=\dfrac{1}{t+2}$;
6) 由式 (5) 计算 $\boldsymbol{w}_{t+1}$;
7) end for;
8) 输出 $\boldsymbol{w}_T$。

因为样本点间满足独立同分布, 所以计算得到的随机次梯度 $\nabla f_i(\boldsymbol{w}_t)$ 是损失函数在点 $\boldsymbol{w}_t$ 处次梯度 $\boldsymbol{g}_t$ 的无偏估计。应用文献 [21] 中将批处理算法收速率转化为随机算法收敛速率的技巧, 可以得到定理 2。

定理 2　设 $f(\boldsymbol{w})$ 是凸函数, $\alpha_t=\alpha/\sqrt{t}$, 取 $\beta_{1,t}=\dfrac{t\sqrt{t}}{(t+2)\sqrt{t-1}}\hat{\boldsymbol{V}}_t^{1/2}\hat{\boldsymbol{V}}_{t-1}^{-1/2}$, $\beta'_{1,t}=\dfrac{1}{t+2}$, $\boldsymbol{w}_t$ 由随机 Adam 算法生成, 则

$$E[f(\boldsymbol{w}_t)-f(\boldsymbol{w}^*)]\leqslant\frac{f(\boldsymbol{w}_0)-f(\boldsymbol{w}^*)}{1+t}+$$

$$\frac{D_\infty^2\sqrt{t}}{2\alpha(1+t)}\sum_{i=1}^{d}\hat{\boldsymbol{V}}_{t,i}^{1/2}+\frac{\alpha G_\infty}{(1+t)\sqrt{1-\beta_2}}\sum_{i=1}^{d}\left\|\boldsymbol{g}_{1:t,i}\right\|_2$$

由定理 2 可知, 得到了随机 Adam 算法在非光滑条件下的最优个体收敛速率。

3　实验

本节主要对随机 Adam 算法的个体收敛速率的理论分析和稀疏性进行实验验证。

3.1　实验数据和比较算法

实验采用 6 个标准数据集, 分别是 ijcnn1、covtype、a9a、w8a、CCAT 和 astro-physic。这些数据均来自于 LIBSVM 网站, 具体描述可见表 1。

表 1　标准数据集描述

Table 1　Introduction of standard datasets

数据集	训练样本数	维数	稀疏度/%
ijcnn1	49 990	22	59.09
covtype	522 911	54	22.12
a9a	24 703	123	11.27
w8a	49 749	300	3.88
CCAT	23 149	47 236	0.16
astro-physic	29 882	99 757	0.08

实验对 5 种随机优化方法进行比较，分别是平均形式输出的 SGD 方法、个体形式输出的 Heavy-ball 型动量法、NAG 型动量法、平均形式输出的 Adam 算法及个体形式输出的 Adam 算法。从理论分析的角度来说，上述 5 种算法的收敛界均达到了最优。

3.2　实验方法及结论

为了算法比较的公平，各个算法在对应数据集上运行 10 次，每次迭代 10 000 步，最后取平均值作为输出。SGD 算法的计算方式为式 (2)，其中步长 $\alpha_t = 1/\sqrt{t}$。Heavy-ball 型动量法的计算方式及步长选取同文献 [11]。NAG 型动量法的计算方式及步长选取同文献 [10]。根据文献 [15]，平均形式输出的 Adam 算法步长 $\alpha_t = 0.1/\sqrt{t}\beta_1 = 0.9, \beta_2 = 0.99$。个体形式输出的 Adam 算法步长设置与迭代次数有关，其中 $\alpha_t = 0.1/\sqrt{t}\beta_2 = 0.99$。本次实验我们调用 SLEP 工具箱来实现投影计算，集合 Q 为 l_1 范数球 $\{\boldsymbol{w}: \|\boldsymbol{w}\|_1 < z\}$，根据数据集的不同，$z$ 的取值也会有相应变化。

图 1 为 5 种算法的收敛速率对比图，纵坐标表示当前目标函数值与最优目标函数值之差。其中绿色、蓝色、青色、粉色、红色分别代表平均形式输出的 SGD 算法、个体形式输出的 Heavy-ball 型动量法、个体形式输出的 NAG 型动量法、平均形式输出的 Adam 算法、个体形式输出的 Adam 算法。可以看到，在 5 000 步迭代之后，5 种算法在 6 个标准数据集上都达到了 10^{-2} 的精度，在迭代 10 000 步后，5 种算法在 6 个标准数据集上都达到了 10^{-4} 的精度。5 种算法的收敛趋势基本相同，这与理论分析基本吻合。

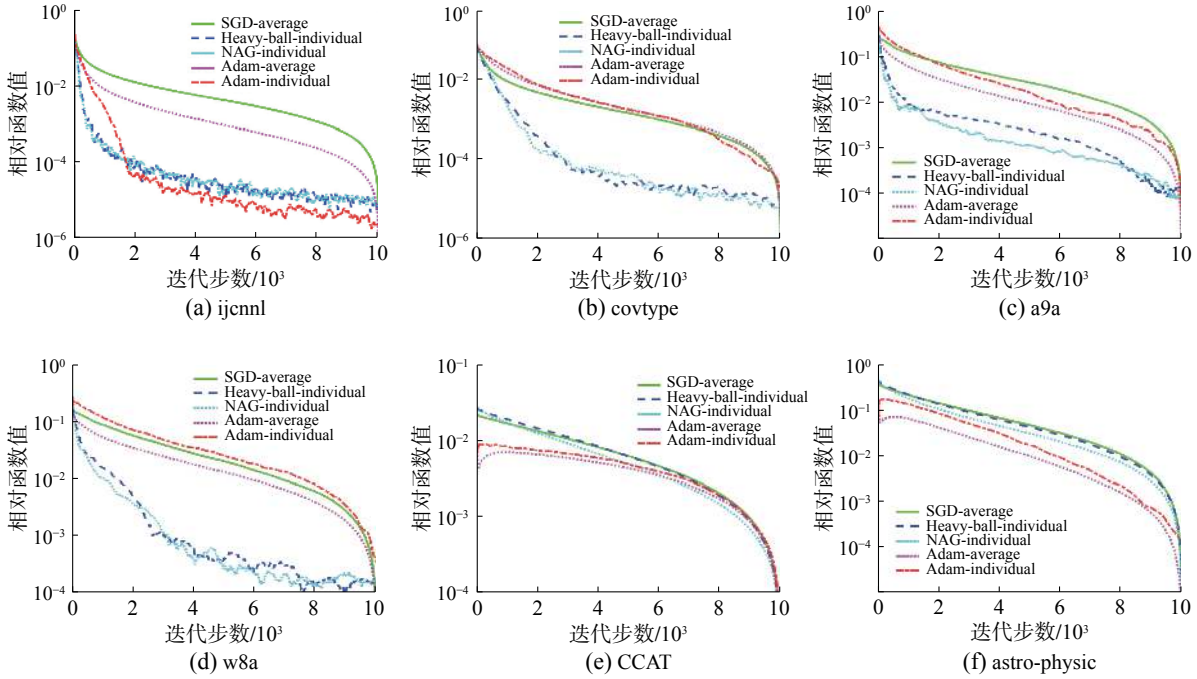

图 1　收敛速率对比

Fig. 1　Comparison of convergence rates

图 2 为 5 种算法的稀疏性对比，纵坐标表示各算法对应输出的稀疏度，稀疏度越低，变量中非零向量所占比例越小。可以看到，个体形式输出的 Heavy-ball 型动量法、NAG 型动量法和 Adam 算法明显比平均形式输出的 SGD 算法和 Adam 算法拥有更低的稀疏度，同时，数据集越稀疏，算法获得的稀疏度也越低。这一结论充分说明，个体解输出比平均解输出能更好地描述样本的稀疏性。同时我们观察到在稀疏度一般的前 4 个数据集上有震荡的现象，这是算法的随机性导致的，在维度较大、稀疏度较低的后两个数据集上该震荡现象消失。

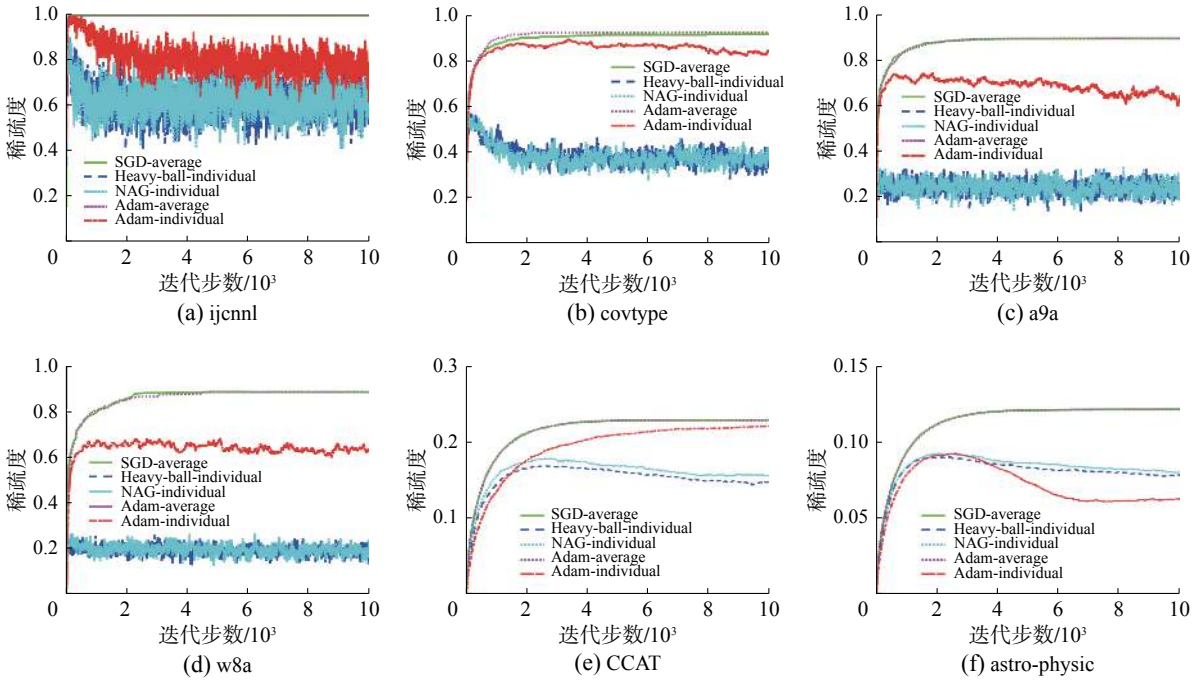

图 2　稀疏性对比

Fig. 2　Comparison of sparsity

4　结束语

本文对非光滑条件下 Adam 算法的收敛性进行了初步的研究, 证明了 Adam 算法可以获得最优的个体收敛速率。本文的结论表明, 在最坏情况下 Adam 算法的个体收敛速率和 Heavy-ball 型动量法的个体收敛速率类似。但 Adam 算法保留了 AdaGrad 算法的优点, 个体收敛速率界比 Heavy-ball 型动量法更紧。下一步将继续研究强凸情况下 Adam 算法最优的个体收敛速率问题以及基于 NAG 型动量的 Adam 算法的最优个体收敛速率问题。

参考文献:

[1] KINGMA D P, BA J L. Adam: a method for stochastic optimization[C]//Proceedings of the 3rd International Conference for Learning Representations. San Diego, USA, 2015.

[2] DUCHI J, HAZAN E, SINGER Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. The journal of machine learning research, 2011, 12: 2121–2159.

[3] ZINKEVICH M. Online convex programming and generalized infinitesimal gradient ascent[C]//Proceedings of the 20th International Conference on Machine Learning. Washington, USA, 2003: 928−935.

[4] TIELEMAN T, HINTON G. Lecture 6.5-RMSProp: divide the gradient by a running average of its recent magnitude[R]. Toronto: University of Toronto, 2012.

[5] ZEILER M D. ADADELTA: an adaptive learning rate method[EB/OL]. (2012–12–22)[2020–04–20]. https://arxiv.org/abs/1212.5701

[6] POLYAK B T. Some methods of speeding up the convergence of iteration methods[J]. USSR computational mathematics and mathematical physics, 1964, 4(5): 1–17.

[7] NESTEROV Y E. A method of solving a convex programming problem with convergence rate $O(1/k^2)$[J]. Soviet mathematics doklady, 1983, 27(2): 372–376.

[8] GHADIMI E, FEYZMAHDAVIAN H R, JOHANSSON M. Global convergence of the Heavy-ball method for convex optimization[C]//Proceedings of 2015 European Control Conference. Linz, Austria, 2015: 310−315.

[9] SHAMIR O, ZHANG Tong. Stochastic gradient descent for non-smooth optimization: convergence results and optimal averaging schemes[C]//Proceedings of the 30th International Conference on International Conference on Machine Learning. Atlanta, USA, 2013: 1-71−1-79.

[10] 陶蔚, 潘志松, 储德军, 等. 使用 Nesterov 步长策略投影次梯度方法的个体收敛性 [J]. 计算机学报, 2018, 41(1): 164–176.
TAO Wei, PAN Zhisong, CHU Dejun, et al. The individual convergence of projected subgradient methods using the Nesterov's step-size strategy[J]. Chinese journal of computers, 2018, 41(1): 164–176.

[11] TAO Wei, PAN Zhisong, WU Gaowei, et al. The strength of Nesterov's extrapolation in the individual convergence of nonsmooth optimization[J]. IEEE transactions on neural networks and learning systems, 2020, 31(7):

2557–2568.

[12] 程禹嘉, 陶蔚, 刘宇翔, 等. Heavy-Ball 型动量方法的最优个体收敛速率 [J]. 计算机研究与发展, 2019, 56(8): 1686–1694.
CHENG Yujia, TAO Wei, LIU Yuxiang, et al. Optimal individual convergence rate of the Heavy-ball-based momentum methods[J]. Journal of computer research and development, 2019, 56(8): 1686–1694.

[13] KIROS R, ZEMEL R S, SALAKHUTDINOV R, et al. A multiplicative model for learning distributed text-based attribute representations[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 2348−2356.

[14] BAHAR P, ALKHOULI T, PETER J T, et al. Empirical investigation of optimization algorithms in neural machine translation[J]. The Prague bulletin of mathematical linguistics, 2017, 108(1): 13–25.

[15] REDDI S J, KALE S, KUMAR S. On the convergence of Adam and beyond[C]//Processing of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018.

[16] WANG Guanghui, LU Shiyin, TU Weiwei, et al. SAdam: a variant of Adam for strongly convex functions[C]//Processing of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020.

[17] CHEN Xiangyi, LIU Sijia, SUN Ruoyu, et al. On the convergence of a class of Adam-type algorithms for non-convex optimization[C]//Processing of the 7th International Conference on Learning Representations. New Orleans, USA, 2019.

[18] DUCHI J, SHALEV-SHWARTZ S, SINGER Y, et al. Efficient projections onto the *l*1-ball for learning in high dimensions[C]//Processing of the 25th International Conference on Machine learning. Helsinki, Finland, 2008: 272−279.

[19] AGARWAL A, BARTLETT P L, RAVIKUMAR P, et al. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization[J]. IEEE transactions on information theory, 2012, 58(5): 3235–3249.

[20] 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.

[21] RAKHLIN A, SHAMIR O, SRIDHARAN K. Making gradient descent optimal for strongly convex stochastic optimization[C]//Processing of the 29th International Conference on International Conference on Machine Learning. Edinburgh, Scotland, 2012: 1571−1578.

作者简介：

黄鉴之，硕士研究生，主要研究方向为凸优化算法及其在机器学习中的应用。

丁成诚，硕士研究生，主要研究方向为凸优化算法及其在机器学习中的应用。

陶卿，教授，博士，主要研究方向为模式识别、机器学习和应用数学。承担国家自然科学基金、安徽省自然科学基金等。发表学术论文 60 余篇。