计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2020.20190462
57(9): 2001-2008, 2020

基于动态约束自适应方法抵御高维鞍点攻击

李德权[1] 许 月[1] 薛 生[2]
[1](安徽理工大学数学与大数据学院 安徽淮南 232001)
[2](安徽理工大学能源与安全学院 安徽淮南 232001)
(dqli@aust.edu.cn)

Defending Against Dimensional Saddle Point Attack Based on Adaptive Method with Dynamic Bound

Li Dequan[1], Xu Yue[1], and Xue Sheng[2]
[1](*School of Mathematics and Big Data, Anhui University of Science and Technology, Huainan, Anhui* 232001)
[2](*School of Energy and Security, Anhui University of Science and Technology, Huainan, Anhui* 232001)

Abstract With the advent of the era of big data, distributed machine learning has been widely applied to process massive data. The most commonly used one is the distributed stochastic gradient descent algorithm, but it is vulnerable to different types of Byzantine attacks. In order to maximize the elastic limit to defend against attacks and optimize objective function in the distributed dimensional Byzantine environment based on the gradient update rule, firstly a new Byzantine attack method—saddle point attack is proposed in this paper. Contrasting with the adaptive non-adaptive methods, the adaptation with dynamic bound escapes the saddle point fast when the objective function is stuck in the saddle point. The comparative experiment is made on the classification of data sets. Secondly, an aggregation rule *Saddle*(·) for filtering Byzantine agents is proposed, and it is proved that the rule is the dimensional Byzantine resilience. Therefore, in the distributed dimensional Byzantine environment, the adaptive optimization method with dynamic bound combined with the aggregation rule *Saddle*(·) can effectively defend against the saddle point attack. Finally, the error rate of the data set classification in the experimental results is compared to analyze the advantages and disadvantages of the adaptation with dynamic bound over the adaptive and non-adaptive methods. The result shows that the adaptation with dynamic bound combined with the aggregation rule *Saddle*(·) is less affected by the saddle point attack in the distributed dimensional Byzantine environment.

Key words distributed optimization; dimensional Byzantine; saddle point attack; the adaption with dynamic bound; the aggregation rule *Saddle*(·)

摘 要 随着大数据时代的到来,分布式机器学习已广泛应用于处理海量数据.其中最常用的是分布式随机梯度下降算法,但其易受到不同类型的 Byzantine 攻击.为了解决在分布式高维 Byzantine 环境下,能最大弹性限度地抵御蓄意攻击问题并有效求解优化问题.基于梯度更新规则,首先提出了一种新的

收稿日期:2019-07-01;**修回日期**:2020-04-03
基金项目:国家重点研发计划项目(2018YFF0301000);国家自然科学基金项目(61472003);安徽省学术和技术带头人及后备人选项目(2019H211);安徽省淮南市"50·科技之星"创新团队项目
This work was supported by the National Key Research and Development Program of China (2018YFF0301000), the National Natural Science Foundation of China (61472003), the Academic and Technical Leaders and the Backup Candidates of Anhui Province (2019H211), and the Program of the Innovation Team of "50 Star of Science and Technology" of Huainan, Anhui Province.
通信作者:许月(xyxuyue7788@163.com)

Byzantine 攻击方式——鞍点攻击.并分析了当目标函数陷入鞍点时,相比较于自适应和非自适应方法,所提出的动态约束自适应方法能够更快逃离鞍点,进而在数据集分类问题上做了比对实验.其次,提出了一种过滤 Byzantine 个体的聚合规则 $Saddle(\cdot)$,理论分析表明它是高维 Byzantine 弹性.因此,在分布式高维 Byzantine 环境下,采用动态约束的自适应优化方法结合聚合规则 $Saddle(\cdot)$能够有效抵御鞍点攻击.最后,从数据集分类实验结果的错误率和误差方面比较并分析了动态约束自适应与自适应和非自适应方法的优劣性.结果表明,结合聚合规则 $Saddle(\cdot)$的动态约束自适应在分布式高维 Byzantine 环境下受鞍点攻击的影响较小.

关键词 分布式优化;高维 Byzantine;鞍点攻击;动态约束自适应;聚合规则 $Saddle(\cdot)$

中图法分类号 TP301.6

随着传感器和智能设备技术的快速发展以及在实际中的广泛运用,协同收集来自多个信息源的大量数据成为现实.面对具有空间分布的海量数据的处理,作为能够从中提取有价值信息的主要有效工具,机器学习在计算机视觉、医疗保健和金融市场分析等广泛领域的应用取得了巨大成功.其中分布式随机梯度下降算法(stochastic gradient descent, SGD)作为最常用的机器学习方法之一,在每步迭代中只随机训练一个样本数据,大大减少了训练时间,但容易受到 Byzantine 攻击[1]的影响.由于系统故障、软件 bug[2]、数据损坏和通信延迟等原因造成的个体或机器之间共享信息错误[3],统称为 Byzantine 攻击.所以,在 Byzantine 环境下研究分布式机器学习如何高效处理共享信息的数据是极为重要的.

目前关于 Byzantine 攻击的相关研究正如火如荼,上述是几种导致故障的实际原因,体现在算法中又分为 Byzantine 个体对共享模型参数 $\widetilde{w}_t^j$[1]和梯度 $\widetilde{g}_i$[4]进行更改.文献[1]在假设 Byzantine 个体数量的上界已知和未知的 2 种情况下,讨论了 Byzantine 个体对共享的参数 $\widetilde{w}_t^j$ 进行的更改:

$$\widetilde{w}_t^j=\begin{cases}w_t^j, & j\in[m]\backslash B,\\ \text{任意值}, & j\in B,\end{cases}$$

其中,B 表示 Byzantine 个体,m 表示总个体数目,$[m]\backslash B$ 表示非 Byzantine 个体.并且将 Byzantine 攻击方式分为 3 种类型:第 1 种“随机”攻击,输出并共享一个随机向量服从(0,1)的均匀分布,即 $\sigma\sim U[0,1]$;第 2 种“反”攻击,共享与模型参数互为相反数的值;第 3 种“噪声”攻击,给模型参数添加一个高斯噪声服从(0,0.01)的正态分布,即 $\varepsilon\sim N[0,0.01]$.文献[2]则假定 Byzantine 个体对共享的梯度 $\widetilde{g}_i$ 进行了如下更改:

$$\widetilde{g}_i=\begin{cases}g_i, & i\in[m]\backslash B,\\ \text{任意值}, & i\in B,\end{cases}$$

进而通过在模型参数 w 上加上一定次数和不同大小的扰动 p,即共享模型参数为 $\widetilde{w}\leftarrow w+p$,能够让随机梯度快速逃离平坦区域,从而逃离了鞍点,这意味着能够有效抵御 Byzantine 攻击.

Byzantine 弹性指的是存在 Byzantine 攻击的情况下算法仍能正常执行的弹性区间[5].文献[6]分析了经典 Byzantine 与高维 Byzantine 的区别与联系,如图 1 所示.文献[6]也将 Byzantine 攻击类型分成了 3 种:高斯攻击、全知型攻击和位翻转型攻击.但仅仅用了 SGD 去解决上述 3 种类型的 Byzantine 攻击的优化问题.

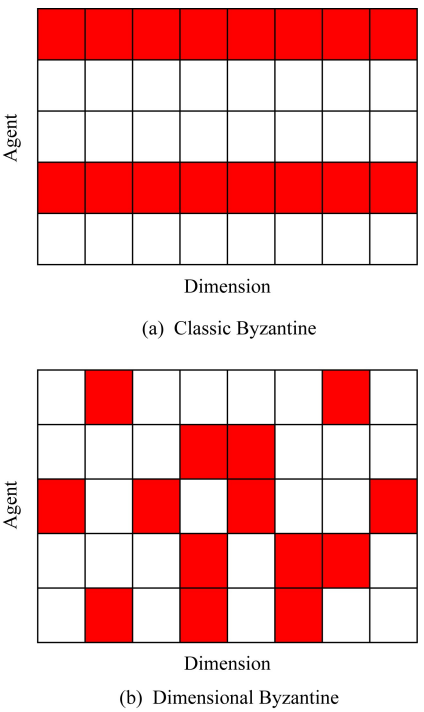


Fig. 1 Two types of Byzantine

图 1 Byzantine 的 2 种不同方式

基于梯度的优化算法会因为鞍点处的梯度值为 0 而易陷入非最优的情形,尤其在高维情况下,包围

鞍点的水平或平坦区域范围较大,这导致分布式SGD很难快速逃离该区域,因而长时间陷于鞍点附近.为了有效解决这一重要问题,受文献[4,6]的启发,提出了一种新的Byzantine攻击类型——高维鞍点攻击.虽与文献[4]一样对共享的梯度 $\tilde{g}_i$ 进行了更改,但不同之处在于梯度更改的值不同,而且是在高维情况下分析的Byzantine攻击,并结合了文献[6]所提出的动态约束自适应优化算法而非传统的SGD.

文献[7]提出用自适应方法代替传统的SGD来有效解决陷入鞍点而无法达到最优值的困境.通过调整随机梯度噪声的方向,使得噪声能在鞍点处是等方向的,有助于朝正确的方向快速逃离鞍点.但自适应方法则面临两大问题:1)与SGD相比泛化能力差;2)由于不稳定或者极端的学习率会导致算法不收敛.文献[8]不仅验证了极端学习率会导致算法性能差,而且提出了给学习率加上动态约束实现从自适应方法到SGD平缓过渡的方法——具有动态约束的自适应矩估计(adaptive moment estimation with dynamic bound, ADABOUND).结果表明这2种动态约束自适应方法消除了自适应方法和SGD的泛化差异,并且提高了学习速度.

为了解决在分布式高维Byzantine环境下,能最大弹性限度抵御攻击问题,从而高效求解优化问题.仿真分析了当目标函数陷入鞍点时,动态约束自适应相比较于自适应和非自适应方法逃离鞍点要更快,并在数据集分类问题上同样进行了比对实验.其次,提出了一种过滤Byzantine个体的聚合规则 $Saddle(\cdot)$,理论证明了它是高维Byzantine弹性的.因此,在分布式高维Byzantine环境下,采用动态约束的自适应优化方法结合聚合规则 $Saddle(\cdot)$ 能够有效避免鞍点攻击.

1 预备知识

定义1. Jensen不等式[9].若对于任意 $\boldsymbol{x}=(x_i)$,参数 $\lambda_i \geqslant 0$ 且 $\sum_i \lambda_i = 1$,凸函数 $f(\boldsymbol{x})$ 满足:$f(\sum_{i=1}^{M}\lambda_i x_i) \leqslant \sum_{i=1}^{M}\lambda_i f(x_i)$.上述公式被称为Jensen不等式.

定义2. 鞍点.解决非线性规划问题 $\min f(\boldsymbol{x})$, s.t. $h_i(\boldsymbol{x}) \geqslant 0, (i=1,2,\cdots,p)$ 的拉格朗日函数,指的是目标函数 $f(\boldsymbol{x})$ 和约束函数 $h_i(\boldsymbol{x})$ 的线性组合,拉格朗日函数为 $L(\boldsymbol{x},\boldsymbol{\lambda}) = f(\boldsymbol{x}) + \sum_{i=1}^{p}\lambda_i h_i(\boldsymbol{x})$,其中,$\boldsymbol{x} \in \mathbb{R}^n$,拉格朗日系数 $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\cdots,\lambda_p)^{\mathrm{T}}$,$\lambda_i \geqslant 0$,$L(\boldsymbol{x}^*,\boldsymbol{\lambda})$ 表示当 $\boldsymbol{x}$ 为固定值 $\boldsymbol{x}^*$ 时,拉格朗日函数为 $L(\boldsymbol{x}^*,\boldsymbol{\lambda})$;同理,当 $\boldsymbol{\lambda}$ 为固定值 $\boldsymbol{\lambda}^*$ 时,拉格朗日函数为 $L(\boldsymbol{x},\boldsymbol{\lambda}^*)$;当有一组解为 $(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$ 时,拉格朗日函数的值为 $L(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$.若 $L(\boldsymbol{x}^*,\boldsymbol{\lambda}) \leqslant L(\boldsymbol{x}^*,\boldsymbol{\lambda}^*) \leqslant L(\boldsymbol{x},\boldsymbol{\lambda}^*)$,则点 $(\boldsymbol{x}^*,\boldsymbol{\lambda}^*)$ 称为鞍点[10].

判断鞍点的一个充分条件是:函数在一阶导数为0处的Hessian矩阵为不定矩阵.

定义3. 严格鞍函数.对于所有的 $\boldsymbol{x}$,满足下列条件:

1) 梯度 $\nabla f(\boldsymbol{x})$ 很大;

2) Hessian矩阵 $\nabla^2 f(\boldsymbol{x})$ 具有负的特征值;

3) 点 $\boldsymbol{x}$ 位于局部极小值附近;

则函数 $f(\boldsymbol{x})$ 是严格鞍函数[11-12].

定义4. 高维Byzantine弹性[6].令 α 是 $[0,\frac{\pi}{2})$ 的任意角,q 是 $[0,m]$ 的任意整数.$(g_i: i\in[m])$ 是在 d 维空间上独立同分布的随机向量,梯度 $\boldsymbol{G}=\nabla f(\boldsymbol{x})$ 满足 $E[\boldsymbol{G}]=\boldsymbol{g}$,其中 $E[\boldsymbol{G}]$ 表示 $\boldsymbol{G}$ 的均值.对于每一维,在这 m 个值中最高有 q 个被任意值取代,即对于第 $j\in[d]$ 维,$\{(\tilde{g}_i)_j: i\in[m]\}$ 有 q 个元素是Byzantine个体所共享的梯度,其中 $(\tilde{g}_i)_j$ 表示向量 $\tilde{\boldsymbol{g}}$ 的第 j 维.高维Byzantine弹性的聚合规则 $Aggr(\cdot)$ 需满足如下条件:

1) $\langle E[Aggr],\boldsymbol{g}\rangle \geqslant (1-\sin\alpha)\|\boldsymbol{g}\|^2 > 0$,其中 $\langle\cdot\rangle$ 表示内积运算.

2) 对于指数 $r=2,3,\cdots,r_1+r_2+\cdots+r_{n-q}=r$,$E[Aggr]^r \leqslant c_1 E[\boldsymbol{G}]^{r_1} + c_2 E[\boldsymbol{G}]^{r_2} + \cdots + c_{n-q}E[\boldsymbol{G}]^{r_{n-q}}$,其中 $c_1,c_2,\cdots,c_{n-q}$ 为线性组合的相关系数.

聚合规则 $Aggr(\cdot)$ 指的是满足规则 $Aggr(\cdot)$ 的被认为是非Byzantine个体共享的真实信息,相反则被认为是Byzantine个体共享的虚假信息,需要被过滤掉.

Reddi等人[13]提出深度学习[14]优化方法的一般框架(包括自适应和非自适应方法):

① $\boldsymbol{g}|_t = (g_1,g_2,\cdots,g_n)|_t = \nabla f(\boldsymbol{x}|_t)$;

② $\boldsymbol{m}|_t = \varphi_t((g_1,g_2,\cdots,g_n)_t)$ 和 $\boldsymbol{v}|_t = \Phi_t((g_1,g_2,\cdots,g_n)_t)$;

③ $\alpha|_t = \dfrac{\alpha}{\sqrt{t}}$;

④ $\boldsymbol{x}|_{t+1} = \boldsymbol{x}|_t - \dfrac{\alpha|_t \boldsymbol{m}|_t}{\sqrt{\boldsymbol{v}|_t}}$;

其中,$\boldsymbol{g}|_t$ 为 $f(\boldsymbol{x}|_t)$ 梯度,α 为深度学习步长,t 为

迭代时间,$\boldsymbol{x}|_t$ 为第 t 步迭代的解$(x_1,x_2,\cdots,x_n)_t$,β_1 和 β_2 分别对应一阶梯度和二阶梯度移动平均的指数,φ_t 和 Φ_t 在自适应和非自适应中分别对应如表 1 所示的值.

Table 1 Adaptive and Non-adaptive Corresponding Values

表 1 自适应和非自适应对应值

Gradient Function	SGD	ADAGRAD	RMSPROP	ADAM
φ_t	$\boldsymbol{g}\|_t$	$\boldsymbol{g}\|_t$	$\boldsymbol{g}\|_t$	$(1-\beta_1)\sum_{i=1}^{t}\beta_1^{t-i}\boldsymbol{g}\|_i$
Φ_t		$\mathrm{diag}(\sum_{i=1}^{t}\boldsymbol{g}\|_i^2)/t$	$(1-\beta_2)\,\mathrm{diag}(\sum_{i=1}^{t}\beta_2^{t-i}\boldsymbol{g}\|_i^2)$	$(1-\beta_2)\,\mathrm{diag}(\sum_{i=1}^{t}\beta_2^{t-i}\boldsymbol{g}\|_i^2)$

2 优化方法与模型建立

2.1 动态约束自适应方法

梯度消失或梯度爆炸是深度学习经常会遇到的难题.如果梯度值很小,就会出现梯度消失的情况,从而导致目标函数长时间收敛不到最小值;若梯度值很大,就会出现函数值上下波动,收敛不到最优值.如下列函数,无论步长大小如何,自适应动量估计算法(adaptive moment estimation, ADAM)都无法收敛到最优值.对于函数序列$\{f_t(z)\}_{t=1}^{T}=[-2,2]$,

$$f_t(z)=\begin{cases}-100, & z<0,\\ 0, & z>1,\\ -z, & 0\leqslant z\leqslant 1, t \bmod C=1,\\ 2z, & 0\leqslant z\leqslant 1, t \bmod C=2,\\ 0, & \text{otherwise},\end{cases}$$

其中 $C\in\mathbb{N}_+$ 并满足 $5\beta_2^{C-2}\leqslant\dfrac{1-\beta_2}{4-\beta_2}$.这个函数序列:当满足 $z<0$ 的任一点,能够得到最小的误差界;假设 $\beta_1=0$,当满足 $z\geqslant 0$ 时,ADAM 收敛到次优解.因为算法每迭代到第 C 步,得到一个梯度值为 -1,导致算法朝着反方向移动.而且在下一步的时候,即第 $C+1$ 步,梯度值为 2.根据表 1 的 ADAM 的梯度更新规则,梯度值越大,对学习率的影响越小.所以第 $C+1$ 步并不能够缓解上一步的错误.综上所述,自适应方法 ADAM 由于梯度值大小的问题而导致不收敛[8].

通过上述分析,自适应方法易受极端学习率影响.为有效克服这一困难,动态约束自适应的算法思想是[8]:通过给梯度加一个阈值区间,并且上下阈值都随时间变化最后收敛到 SGD 的学习率,从而实现自适应方法向 SGD 的平缓过渡.

针对梯度消失和梯度爆炸问题,不妨假设 $g_1=\dfrac{\partial f(w)}{\partial w_1}$,$g_2=\dfrac{\partial f(w)}{\partial w_2}$,$\cdots$,$g_m=\dfrac{\partial f(w)}{\partial w_m}$,其中 $\|\boldsymbol{g}\|_2=\sqrt{g_1^2+g_2^2+\cdots+g_m^2}$,并设定裁剪阈值为 $\eta_{\max}(t)$ 和 $\eta_{\min}(t)$.

当 $\|\boldsymbol{g}\|_2\geqslant\eta_{\max}(t)$ 时,$\boldsymbol{g}=\dfrac{\eta_{\max}(t)}{\|\boldsymbol{g}\|_2}\cdot\boldsymbol{g}$;

当 $\|\boldsymbol{g}\|_2<\eta_{\min}(t)$ 时,$\boldsymbol{g}=\dfrac{\eta_{\min}(t)}{\|\boldsymbol{g}\|_2}\cdot\boldsymbol{g}$.

文献[8]已经证明了动态约束自适应算法收敛,它的遗憾界是 $R_T\leqslant O(\sqrt{T})$.

2.2 鞍点攻击与聚合规则 *Saddle*(·)

定义 5. 鞍点攻击类型为

$$(\tilde{g}_i)_j=\begin{cases}(g_i)_j, & i\in[m]\backslash B,\\ 0, & i\in B,\end{cases}$$

其中,$(g_i)_j$ 表示第 j 维第 i 个个体的梯度值,$j\in[d]$.$i\in B$ 时,同时满足 Hessian 矩阵是不定的.鞍点攻击影响梯度,从而影响学习率,导致目标函数不收敛,所以需要寻求一种聚合规则过滤 Byzantine 个体.下面提出的聚合规则[15]可以过滤高维鞍点个体,实现抵御攻击的目的.

定义 6. 聚合规则 $Saddle(\cdot)$为

$$\boldsymbol{s}=Saddle((\tilde{g}_i:i\in[m])),$$

其中,对任意维 $j\in[d]$,$s_j\in\max_{m-q}\{(\tilde{g}_i)_j+p\}$ 表示给每个梯度值加噪声扰动 p,然后选择最大的 $m-q$ 个作为非 Byzantine 梯度集.因为鞍点性质本身具有不稳定性,当给一个处在鞍点位置的球施加一些轻微的扰动,该球可能会从鞍点处滑落,从而实现逃离鞍点找到严格鞍函数的目的.

定理 1. 噪声梯度下降法能够在多项式时间内找到严格鞍函数的局部最小值点[10].

定理 1 验证了聚合规则 $Saddle(\cdot)$是能在有限时间内找到函数最优值,即逃离了鞍点.为了验证聚合规则 $Saddle(\cdot)$更具有一般性,下面证明它是高维 Byzantine 弹性的.

3 聚合规则 *Saddle*(·)

定理 2. 令$(g_i:i\in[m])$是在 d 维空间上独立同分布的随机向量,其中 $E[\boldsymbol{G}]=\boldsymbol{g}$,假设 $E[G_i-g_i]^2=\sigma^2$,所以 $E\|\boldsymbol{G}-\boldsymbol{g}\|^2=E\sum_{j=1}^{d}[(G_i)_j-(g_i)_j]^2=\sum_{j=1}^{d}\sigma^2=d\sigma^2$. 对于任意维 $j\in[d]$,在这 m 个值中有 q 个被任意值取代,其中$(\tilde{g}_i)_j$ 表示向量 $\tilde{\boldsymbol{g}}_i$ 的第 j 维.如果 $q\leqslant\left[\frac{m}{2}\right]-1$,$\eta(m-q)\sqrt{d}\sigma<\|\boldsymbol{g}\|$,其中 $\eta(m-q)=\sqrt{m-q}$,那么聚合规则 *Saddle*(·)是高维 Byzantine 弹性.

证明.

$$E[s_j-g_j]^2\leqslant E\left[\max_{i\in[m]\backslash B}((\tilde{g}_i)_j-g_j)^2\right]\leqslant E\left[\left(\min_{i\in[m]\backslash B}((\tilde{g}_i)_j+p)-g_j\right)^2\right]\leqslant E\left[\sum_{i\in[m]\backslash B}((\tilde{g}_i)_j-g_j)^2\right]=\sum_{i\in[m]\backslash B}E[((\tilde{g}_i)_j-g_j)^2]=(m-q)E[G_j-g_j]^2=(m-q)\sigma_j^2.$$

根据 Jensen 不等式,

$$\|E[s_j]-g_j\|^2\leqslant E\|s_j-g_j\|^2=E[\sum_{j=1}^{d}(s_j-g_j)^2]=\sum_{j=1}^{d}E[(s_j-g_j)^2]\leqslant\sum_{j=1}^{d}(m-q)\sigma_j^2=(m-q)\sum_{j=1}^{d}\sigma_j^2=(m-q)d\sigma^2.$$

由假设知,$\eta(m-q)\sqrt{d}\sigma<\|\boldsymbol{g}\|$,即 $E[\boldsymbol{s}]$在一个以 $\boldsymbol{g}$ 为中心、$\eta(m-q)\sqrt{d}\sigma$ 为半径的圆内.那么

$$\langle E[\boldsymbol{s}],\boldsymbol{g}\rangle\geqslant(1-\sin^2\alpha)\|\boldsymbol{g}\|^2\geqslant(1-\sin\alpha)\|\boldsymbol{g}\|^2,$$

其中 $\sin\alpha=\frac{\eta(m-q)\sqrt{d}\sigma}{\|\boldsymbol{g}\|}$.

以上验证了聚合规则 *Saddle*(·)满足高维 Byzantine 弹性的条件 1).

根据等价范数的定义:有限维空间上的任何 2 个范数必是等价的.存在常数 c_1,使得

$$\|\boldsymbol{s}\|=\sqrt{\sum_{j=1}^{d}s_j^2}\leqslant\sqrt{\sum_{j=1}^{d}((s_i)_j+p)^2}\leqslant\sqrt{\sum_{j=1}^{d}\max_{i\in[m]\backslash B}((\tilde{g}_i)_j)^2}\leqslant\sqrt{\sum_{j=1}^{d}\sum_{i\in[m]\backslash B}((\tilde{g}_i)_j)^2}=\sqrt{\sum_{i\in[m]\backslash B}\|\tilde{g}_i\|^2}\leqslant c_1\sum_{i\in[m]\backslash B}\|\tilde{g}_i\|.$$

不失一般性,用$\{g_1,g_2,\cdots,g_{m-q}\}$ 表示$\{(\tilde{g}_i)_j:i\in[m]\backslash B\}$.因此,存在常数 c_2,使得

$$\|\boldsymbol{s}\|^r\leqslant c_2\sum_{r_1+\cdots+r_{m-q}=r}\|g_1\|^{r_1}\cdots\|g_{m-q}\|^{r_{m-q}}$$

$$E\|\boldsymbol{s}\|^r\leqslant c_1E\|g_1\|^{r_1}+c_2E\|g_2\|^{r_2}+\cdots+c_{m-q}E\|g_{m-q}\|^{r_{m-q}},$$

其中,$r_1+r_2+\cdots+r_{m-q}=r$.所以验证了聚合规则 *Saddle*(·)满足高维 Byzantine 弹性的条件 2).

以上证明了聚合规则 *Saddle*(·)是更具有一般性的高维 Byzantine 弹性,即过滤鞍点的聚合规则是泛化的. 证毕.

4 实验结果

4.1 动态约束自适应方法与其他优化算法的比较

以凸函数:

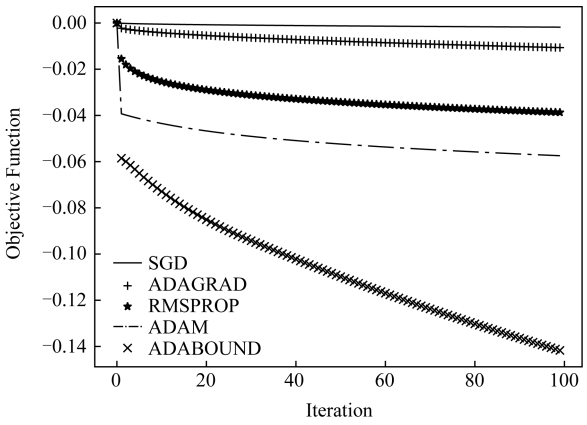
$$f_t(x_1,x_2)=\left[\frac{1}{2}(x_1^2-x_2^2)+b_1x_1+b_2x_2\right]_t$$

为例,其中 b_1,b_2 为随机系数.由鞍点判定条件知,函数 $f_t(x_1,x_2)$的 Hessian 矩阵是不定的,所以该函数存在鞍点.通过不断对(x_1,x_2)进行梯度更新,最终找到目标函数的最优值.因为不同的优化方法对梯度更改规则不同(即学习率不同),所以达到目标函数最优值的快慢不同.

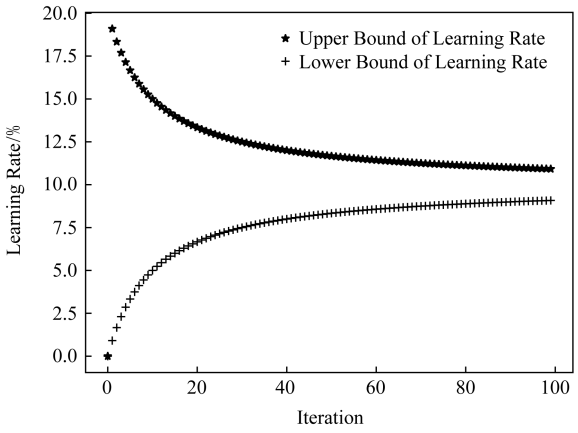
如图 2(a)所示,动态约束自适应(ADABOUND)方法相比较于自适应梯度算法(adaptive gradient, ADAGRAD)、平方根改进算法(root mean square prop, RMSPROP)、ADAM 和 SGD 逃离鞍点的速度明显比较快.图 2(b)表示动态约束自适应的上下阈值随时间变化,最后都收敛到与 SGD 学习率一致,实现了自适应到非自适应的平缓过渡.

不仅如此,在随机生成的(50 000×30)数据集中,分别用不同的优化方法将该数据集分成 10 个类.逻辑回归用以解决该分类问题,对于输入的$\boldsymbol{x}|_t$,其对应的类标签为 $\boldsymbol{y}|_t$,我们的目标是训练分类模型 $\boldsymbol{y}|_t=w\boldsymbol{x}|_t+\boldsymbol{b}$ 的参数 w 和 $\boldsymbol{b}$,使得属于当前类的概率最大.下面实验通过比较错误率(error rate)与误差(error)来分析方法的优劣.

如图 3(a)所示,在数据集分类问题上自适应方法虽然在前期分类的精确度比 SGD 差一点,但随着训练次数的增加,最后自适应方法明显比 SGD 的分类错误率要低的多.图 3(b)显示,动态约束自适应都比自适应的分类错误率要低。通过图 3(c)发现,当前值与真实值的距离测算的误差,动态约束自适应



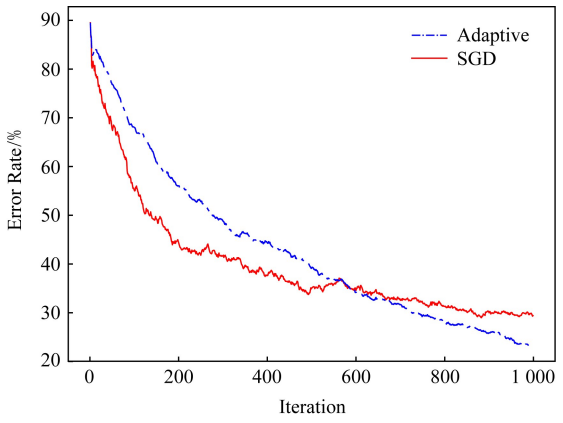
(a) Five optimization methods to escape from saddle point



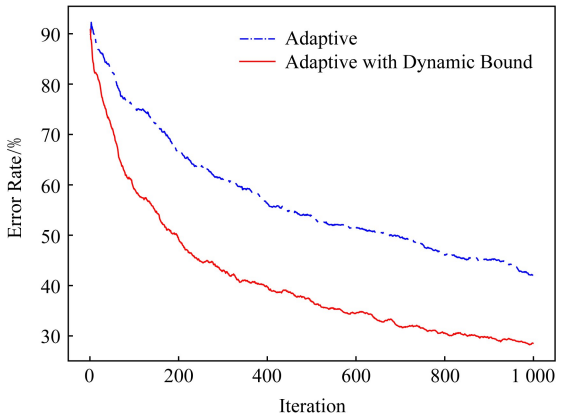
(b) Learning rate of adaptive method with dynamic bound

Fig. 2 Comparison of the optimization methods used to escape from saddle point

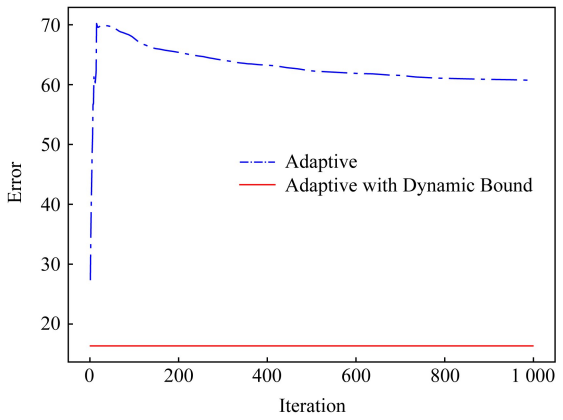
图 2 优化方法用于逃离鞍点的快慢比较



(a) Adaptive method and SGD



(b) Classification error rates of different adaptive methods



(c) Classification error of different adaptive methods

Fig. 3 Performance comparison of data set classification by different optimization methods

图 3 不同优化方法分类的性能比较

明显比自适应要低.因为错误率$=\dfrac{\text{训练值}-\text{实际值}}{\text{训练值}}$,由图 3(c)知,ADABOUND 比 SGD 的误差值小,导致误差率却比 SGD 大的原因是因为在不断训练过程中,训练值越来越接近实际值,并最终趋向一个稳定值.图 3(b)表明在 10 000 次以后,SGD 的错误率仍在向下延伸,说明训练值并没有趋于稳定状态.

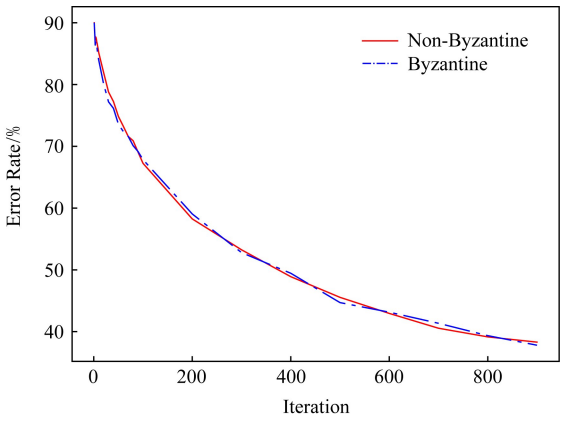
4.2 Byzantine 环境下的动态约束自适应

根据 2.2 节提出的鞍点攻击,该攻击方式会影响梯度,所以对于基于梯度的优化方法就会有一定程度的影响.为了证明动态约束自适应能控制极端梯度的影响,在 Byzantine 和非 Byzantine 环境下做了动态约束自适应受环境影响程度的对比实验.并且与文献[6]作比对实验,观察 SGD 受 Byzantine 攻击的影响大小.

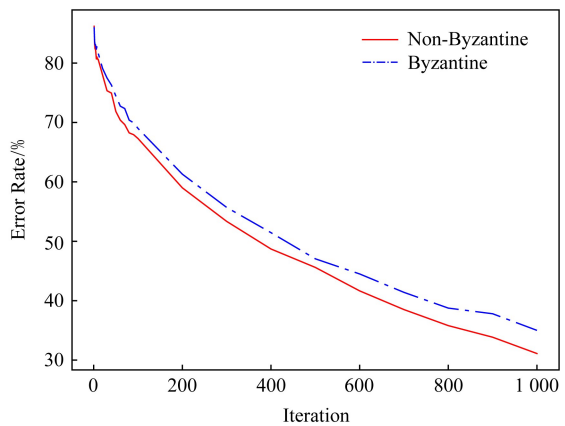
如图 4(a)(b),动态约束自适应方法相比较于 SGD 在 Byzantine 环境下不会受很大的影响.因为经过梯度裁剪的自适应,不会因为梯度的忽大忽小而导致不收敛.图 4(c)表示从误差角度分析,在 Byzantine 和非 Byzantine 环境下动态约束自适应受影响的情况.

4.3 分布式高维 Byzantine 环境下的动态约束自适应

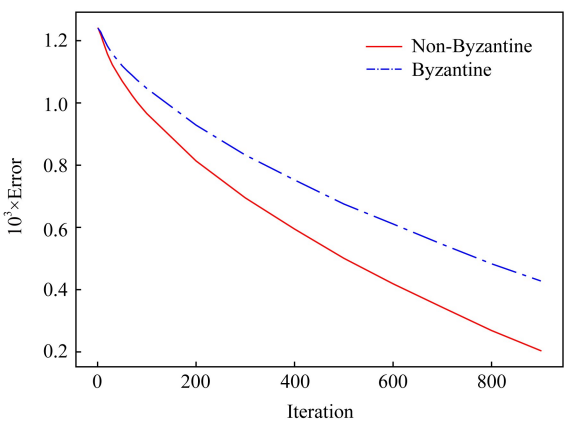
第 3 节已经验证了聚合规则 $Saddle(\cdot)$具有高



(a) Error rate of adaptive with dynamic bound



(b) Error rate of SGD



(c) Error of adaptive with dynamic bound

Fig. 4 Performance comparison of optimization methods in Byzantine and non-Byzantine environments

图 4 Byzantine 和非 Byzantine 环境下的性能比较

维 Byzantine 弹性,所以在分布式环境下鞍点攻击的是任意维的个体.不同个体对应不同的工作机器,分布式计算模型参数,然后它们向主机发送局部模型参数[16].所以在分布式环境下研究鞍点攻击,通过观察不同个体总数目的错误率.

如图 5 所示,不同个体总数目下动态约束自适应方法的分类效果不一样[17].由于高维 Byzantine 在每一维的随机性较大,所以这也是影响图 5 结果的直接原因.

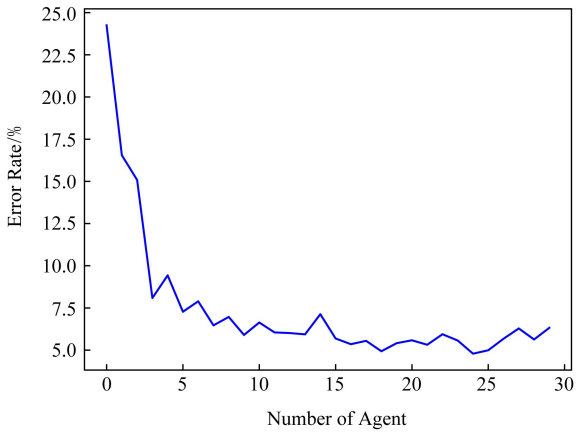


Fig. 5 Error rate corresponding to the total number of different individuals

图 5 不同个体总数所对应的错误率

5 总结与展望

结合文献[8]提出的动态约束自适应并应用于分布式 Byzantine 环境,本文主要提出一种具有高维 Byzantine 弹性的聚合规则 $Saddle(\cdot)$,可有效过滤鞍点攻击的个体,从而实现在确保算法收敛的情况下抵御鞍点攻击的目的.针对数据集分类结果的错误率和误差 2 个重要指标,通过大量数值仿真比较分析动态约束自适应与自适应和非自适应方法的优劣性.结果表明:结合聚合规则 $Saddle(\cdot)$ 的动态约束自适应在分布式 Byzantine 环境下受鞍点攻击的影响较小.

下一步工作希望能够给出不同个体总数所对应不同错误率大小的理论解释,并且能够更大限度地降低错误率.

参 考 文 献

[1] Jin Richeng, He Xiaofan, Dai Huaiyu. Distributed Byzantine tolerant stochastic gradient descent in the era of big data [EB/OL]. 2019 [2019-05-12]. https://arxiv.org/abs/1902.10336v3

[2] Fu Yiqi, Dong Wei, Yin Liangze, et al. Software defect prediction model based on the combination of machine learning algorithms [J]. Journal of Computer Research and Development, 2017, 54(3): 633-641 (in Chinese)

(傅艺绮, 董威, 尹良泽, 等. 基于组合机器学习算法的软件缺陷预测模型[J]. 计算机研究与发展, 2017, 54(3): 633-641)

[3] He Yingzhe, Hu Xingbo, He Jinwen, et al. Privacy and security issues in machine learning systems: A survey [J]. Journal of Computer Research and Development, 2019, 56(10): 2049-2070 (in Chinese)
(何英哲, 胡兴波, 何锦雯, 等. 机器学习系统的隐私和安全问题综述[J]. 计算机研究与发展, 2019, 56(10): 2049-2070)

[4] Yin Dong, Chen Yudong, Ramchandran K, et al. Defending against saddle point attack in Byzantine-robust distributed learning [EB/OL]. 2018[2018-10-12]. https://arxiv.org/abs/1806.05358

[5] Li Junfei, Hu Yuxiang, Wu Jiangxing. Research on improving the control plane's reliability in SDN based on Byzantine fault-tolerance [J]. Journal of Computer Research and Development, 2017, 54(5): 952-960 (in Chinese)
(李军飞, 胡宇翔, 邬江兴. 基于拜占庭容错提高 SDN 控制层可靠性的研究[J]. 计算机研究与发展, 2017, 54(5): 952-960)

[6] Xie Cong, Koyejo O, Gupta I. Generalized Byzantine-tolerant SGD [EB/OL]. 2018 [2019-04-15]. https://arxiv.org/abs/1802.10116

[7] Staib M, Reddi S J, Stayen K, et al. Escaping saddle points with adaptive gradient methods [EB/OL]. 2019 [2019-03-12]. https://arxiv.org/abs/1901.09149

[8] Luo Liangchen, Xiong Yuanhao, Liu Yan, et al. Adaptive gradient methods with dynamic bound of learning rate[EB/OL]. 2019 [2019-03-28]. https://arxiv.org/abs/1902.09843

[9] Jensen J. Sur les fonctions convexes et les inégualités entre les valeurs moyennes [J]. Acta Matem, 1906, 30(1): 175-193

[10] Gu Guoli, Liu Qian, Li Min. Augmented Lagrangian saddle point in semi-infinite planning [J]. Journal of Chongqing Normal University: Natural Science, 2016, 33(6): 13-17 (in Chinese)
(谷国利, 刘茜, 李敏. 半无限规划中的增广拉格朗日鞍点[J]. 重庆师范大学学报: 自然科学版, 2016, 33(6): 13-17)

[11] Ju Sun, Qing Qu, Wright J. When are nonconvex problems not scary [EB/OL]. 2015 [2019-05-31]. https://arxiv.org/abs/1510.06096

[12] Ge Rong, Huang Furong, Jin Chi, et al.Escaping from saddle points-online stochastic gradient for tensor decomposition [C] // Proc of the 28th Conf on Learning Theory. Piscataway, NJ: IEEE, 2015: 797-842

[13] Sashank J, Reddi S J, Stayen K,et al. On the convergence of ADAM and beyond [C/OL] //Proc of the Int Conf on Learning Representations. Montreal, Canada: MILA, 2018 [2019-05-20]. https://openreview.net/forum?id=ryQu7f-RZ

[14] Georgios D, Mahdi M, Rachid G, et al. Aggregathor: Byzantine machine learning via robust gradient aggregation [C/OL] //Proc of the Conf on Systems and Machine Learning. Romandie, Switzerland: EPFL, 2019 [2019-05-20]. https://www.researchgate.net/publication/332947090_AGGREGATHOR_Byzantine_Machine_Learning_via_Robust_Gradient_Aggregation

[15] Lamport L, Shostak R, Pease M. The Byzantine general's problem [J]. ACM Transactions on Programming Languages & Systems, 1982, 4(3): 382-401

[16] Li Xiaoling, Wang Huaimin, Guo Changguo, et al. Research and development of distributed constrained optimization problems [J]. Chinese Journal of Computers, 2015, 38(8): 1656-1671 (in Chinese)
(李小玲, 王怀民, 郭长国, 等. 分布式约束优化问题研究及其进展[J]. 计算机学报, 2015, 38(8): 1656-1671)

[17] Gong Maoguo, Cheng Gang, Jiao Licheng, et al. Nondominated individual selection strategy based on adaptive partition for evolutionary multi-objective optimization [J]. Journal of Computer Research and Development, 2011, 48(4): 545-557 (in Chinese)
(公茂果, 程刚, 焦李成, 等. 基于自适应划分的进化多目标优化非支配个体选择策略[J]. 计算机研究与发展, 2011, 48(4): 545-557)

Li Dequan, born in 1973. PhD, professor. His main research interests include distributed optimization, machine learning and control of sensor network.

Xu Yue, born in 1995. Master candidate. Her main research interests include distributed optimization.

Xue Sheng, born in 1964. Professor, PhD supervisor. His main research interests include coal and gas outburst, gas drainage in goaf and mining area, coal and gas mining and coal spontaneous combustion.