拜占庭共识中基于随机梯度下降的聚合规则

王天翔，郑忠龙*，唐长兵，彭浩，陈中育，吴德，张咪

（浙江师范大学计算机系，浙江　321004）

摘　要：随着数据化时代的变革，区块链技术随着日益剧增的数据量慢慢兴起，区块链技术大大简化了人们的交易程序，并提供了更安全更高效的交易流程.在区块链核心技术的共识算法中，往往会遇到典型的拜占庭问题.对于拜占庭故障，机器学习中常常应用随机梯度下降方法进行处理.本文主要基于随机梯度下降方法提出了截尾均值聚合规则，同时从理论上证明该聚合规则满足拜占庭弹性条件，且具有类线性复杂度，同时实验也证明了拜占庭算法中，基于截尾均值的聚合规则具有强鲁棒性.

关键词: 拜占庭故障；梯度下降法；聚合规则

中国法分类号：

Aggregation rules based on stochastic gradient descent in Byzantine consensus

WANG Tian-Xiang，ZHENG Zhong-Long*，TANG Chang-Bing，PENG Hao，Chen Zhong-Yu，WU De，ZHANG Mi

(Department of computer science，Zhejiang Normal University, Zhejiang 321004,China)

Abstract: With the changes in the era of data, block chain technology has gradually emerged with the ever-increasing amount of data. Block chain technology greatly simplifies people's transaction procedures and provides a safer and more efficient transaction flow. In the consensus algorithm of the core technology of the chain, the typical Byzantine problem is often encountered. For the Byzantine fault, the stochastic gradient descent method is often used in machine learning. In this paper, the censored mean aggregation rule is proposed based on the stochastic gradient descent method. At the same time, it is proved theoretically that the aggregation rule satisfies the Byzantine elastic condition and has linear complexity. At the same time, the experiment also proves that the Byzantine algorithm is based on the truncated mean aggregation rules are robust.

Key words: Byzantine failure; gradient descent method; aggregation rule

机器学习作为人工智能的重要分支已被人们深知并得到了广泛的应用.在人工智能领域，深度学习，人脸识别等都是深深结合机器学习相关知识而开展的[1].在医疗诊断，驾驶导航，股票金融预测，游戏制作领域[5]等，机器学习都有着深远的意义[3].另外，在机器学习中，分布式计算在现代数据密集型应用中变得越来越重要。在许多应用中，大规模数据集分布在多台机器上进行并行处理，以加速计算。

在分布式系统中，我们通常采用状态机副本复制协议来屏蔽一般的故障问题[2].在基于此原理的分布式机器学习中，我们可以将其分为两种方式：（1）多个进程在更新本地参数向量的数据样本上达成一致[4]；（2）多个进程在如何更新参数向量时达成一致.在（1）中数据样本必须传输到每个个体节点中，需要较大的成本，在（2）中，就更新的参数向量，我们无法检测到它是否是真实有效（因为里面可能混有拜占庭向量）.在本文中，我们考虑最常见的故障模型，即拜占庭故障，攻击者可以知道进程的任何信息，并在节点传输信息过程中可以采用任意值发起攻击[6].更具体地说，机器之间的数据传输可以用任意值代替.

在机器学习中，我们通常采用梯度下降方法来优化目标函数[7].在解决拜占庭故障问题时，通常采用随机梯度下降方法（SGD）来优化问题，主要是因为 SGD 在数学期望意义上可以更方便证明其收敛.分布式实现 SGD 通常用以下步骤来表示：参数服务器执行同步回合[8]，在每个回合期间，参数向量被广播给矿工节点.接下来，每个矿工节点计算应用的更新估计（梯度的估计），并且参数服务器汇总其结果以最终更新参数向量.在矿工节点进行参数梯度更新操作时，通常会出现拜占庭故障.

在应对拜占庭矿工节点的攻击时，本文提出了基于截尾均值的聚合规则，且 SGD 在一定条件下任然保持拜占庭弹性：对于每个维度向量，在 n 个矿工节点提供的所有 n 个值中，拜占庭值的数量必须小于 n 的一半. 目前，在处理广义拜占庭 SGD 问题中，已有多种强有力的鲁棒性聚合规则，例如 Geometric Median、Marginal Median、Beyond Median.并且在接受高斯攻击，Bit-flip 攻击等时，依然能够保持鲁棒性。

本文第一节首先介绍了系统模型，简述了 SGD 工作原理，并概括了基于聚合规则的更新迭代法则；第二节给出了拜占庭弹性概念，以及满足拜占庭恢复的聚合规则所必备的条件；第三节概括了基于中值的聚合规则，并给予严格的理论证明；在第四部分，本文采用了轻微的高斯攻击以及全方位的攻击验证在接受外界攻击下，拜占庭聚合规则的鲁棒性。在总结部分，本文对拜占庭式机器学习热点研究方向进行了讨论.

1 系统模型

假设一个分布式系统是由参数向量，n 个矿工节点构成的，在 n 个矿工节点中可能有 f 个拜占庭故障节点，计算可以被分为无限多的同步轮次. 在 t 轮间，参数服务器广播参数向量（$\mathrm{x}_{\mathrm{t}} \in R^{\mathrm{d}}$）至所有的矿工节点，同时每一个正确的矿工节点 p 计算$V_{\mathrm{p}}^{\mathrm{t}}=G\left(\mathrm{x}_{\mathrm{t}},\ \xi_{\mathrm{p}}^{\mathrm{t}}\right)$梯度

$\Delta \mathrm{Q}(\mathrm{x}_{\mathrm{t}})$ 的代价函数Q. 此时，拜占庭矿工节点i提出一个任意向量$V_{\mathrm{i}}^{\mathrm{t}}$，由于通信是同步的，所以在节点反馈信息过程中，如果参数服务器没有收到来自拜占庭节点的信息那就默认为 0.

使用聚合规则[10] $A\mathrm{ggr}(.)$，参数服务器更新节点可以表达如下：

$$\mathrm{x}^{\mathrm{t}+1} \leftarrow \mathrm{x}^{\mathrm{t}}-\gamma^{\mathrm{t}} A\mathrm{ggr}\left(\left\{\widetilde{v}_i^t : i \in [n]\right\}\right) \tag{1}$$

γ^{t}是指学习率. 在没有拜占庭故障情况时，第 i 个矿工节点计算可得 $\mathrm{v}_{\mathrm{i}}^{\mathrm{t}} \sim G^{\mathrm{t}}$，

$G^{t}=\nabla f\left(x^{t},\ \xi\right)$. 当存在拜占庭故障时， v_i^t 可以被任意值所替代，用来替代的就为拜占庭值 $\tilde{v}_i^t$.

2 拜占庭弹性

定义具有鲁棒性聚合规则函数 $A\mathrm{ggr}(.)$，此函数指向优化的最陡方向. 我们用图（1）以及定义 1 来表示满足经典拜占庭弹性的情况：

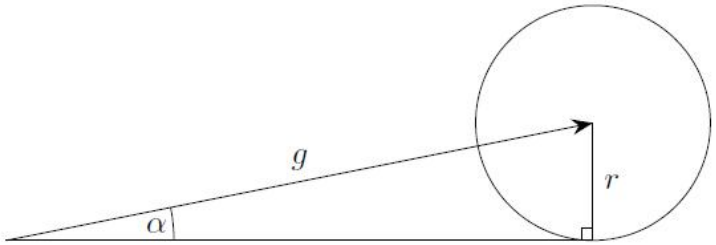

图 1：如果 $\|E[A\mathrm{ggr}]-g\|\le r$，且 $\langle E[A\mathrm{ggr}],\ g\rangle$ 以 $(1-\sin\alpha)\|g\|^2$ 为界，其中 $\sin\alpha=r/\|g\|$.

并且 $A\mathrm{ggr}(.)$ 的矩应该由（正确的）梯度估计器 G 的矩控制[10].

根据 Blanchard et al.[9]我们对经典拜占庭弹性进行如下定义：

定义 1. 假设 $\alpha\in\left(0,\frac{\pi}{2}\right]$，且 $q\in[0,\ n]$. 令 $\{v_i:\ i\in[n]\}$ 在 R^d 上是独立同分布的，且 $v_i\sim G$，$E[G]=g$. 在 $\{v_i:\ i\in[n]\}$ 中，我们假设其中有 q 个向量是任意的（拜占庭式向量）我们记为 $\{\tilde{v}_i:\ i\in[n]\}$. 那么聚合规则 $A\mathrm{ggr}(.)$ 只要满足以下两个条件便是满足拜占庭弹性[11]的：

(1) $\langle E[A\mathrm{ggr}],\ g\rangle\ge(1-\sin\alpha)\|g\|^2>0$；

(2) 当 r=2, 3, 4，$E\|A\mathrm{ggr}\|^r$ 可以被 $E\|G\|^{r_1}$，，$E\|G\|^{r_{n-q}}$ 和 $r_1+\quad+r_{n-q}=r$ 线性组合所限定.

根据 Cong Xie et al.[13]我们对维度拜占庭弹性进行如下定义：

定义 2. 假设 $\alpha\in\left(0,\frac{\pi}{2}\right]$，且 $q\in[0,\ n]$. 令 $\{v_i:\ i\in[n]\}$ 在 R^d 上是独立同分布的，且在 $\{v_i:\ i\in[n]\}$ 中，对于第 j 维向量，假设其中有 q 个向量是任意的（拜占庭式向量）记为 $\{(\tilde{v}_i)_j:\ i\in[n]\}$. 那么聚合规则 $A\mathrm{ggr}(.)$ 只要满足以下两个条件便是满足拜占庭弹性的：

(1) $\langle E[A\mathrm{ggr}],\ g\rangle\ge(1-\sin\alpha)\|g\|^2>0$；

(2) r=2, 3, 4， $E\|Aggr\|^{r}$ 可以被 $E\|G\|^{r1}$， ， $E\|G\|^{r_{n-q}}$ 和 $r_1 + \quad + r_{n-q} = r$ 线性组合

$\langle E[Aggr], g\rangle \geq (1-\sin\alpha)\|g\|^2 > 0$ 所限定.

本文主要讨论在拜占庭环境下，截尾均值的聚合规则可以满足拜占庭弹性条件，并从理论上进行了严格的证明.

3 基于中值的聚合规则

3.1 基于均值的聚合规则

基于均值的聚合规则无法满足拜占庭恢复条件[12]。为证明此结论，我们利用 Blanchard[9] 提出的引理 1 进行证明.

引理 1. 任给一个选择函数 F_{lin}

$$F_{lin}(V_1, \quad , V_n) = \sum_{i=1}^{n} \lambda_i \cdot V_i \tag{2}$$

λ_i 是非零标量. 设 U 是 R^d 中的任一向量，拜占庭矿工节点可以让 F 总是选择 U .

证明：如果拜占庭矿工节点提出向量 $V_n = \frac{1}{\lambda_n} \cdot U - \sum_{i=1}^{n-1} \frac{\lambda_i}{\lambda_n} V_i$ ，且 $F = U$. 参数服务器可以通过设置 λ 来取消拜占庭行为，当 $\lambda_n \to 0$ 它就非拜占庭向量[14]. 从其中我们就可以看出利用平均值来更新迭代向量，并不能满足拜占庭弹性，无法保证

$$\langle E[Aggr], g\rangle \geq (1-\sin\alpha)\|g\|^2 \tag{3}$$

均值仅仅是一组数据的平均，并不具有强鲁棒性，所以本文考虑利用中值来更新聚合规则.

3.2 krum

根据 Blanchard et al.[9]我们对 krum 函数进行以下定义：

$$Krum(\{\tilde{v}_i : i \in [n]\}) = \tilde{v}_k$$
$$k = \arg\min_{i \in [n]} \sum_{i \to j} \|\tilde{v}_i - \tilde{v}_j\|^2 \tag{4}$$

Krum 函数是经典拜占庭[9]的聚合规则，可以表示为：

定理 1. 令 $v_1, \ldots, v_n$ 是任意 d 维独立同分布向量，并且满足 $v_i \sim G$， $E[G] = g$， $E[G-g]^2 = d\delta^2$. $b_1, \ldots, b_q$ 是 $\{v_i : i \in [n]\}$ 中替代原向量的 d 维任意拜占庭向量，如果满足 $2q+2 < n$ 且 $\eta_0(n, q)\sqrt{d}\delta < \|g\|$，

$$\eta_0^2(n, q) = 2\left(n - q + \frac{q(n-q-2) + q^2(n-q-1)}{n-2q-2}\right) \tag{5}$$

Krum 函数满足经典$(\alpha_0,\ \mathrm{q})$-拜占庭弹性，

其中，

$$\alpha_0 \in \left(0,\frac{\pi}{2}\right],\ \sin\alpha_0 = \frac{\eta_0(\mathrm{n},\ \mathrm{q})\sqrt{\mathrm{d}}\delta}{\|\mathrm{g}\|} \tag{6}$$

3.3 几何中值

根据 Cong Xie et al.[13]，我们对几何中值聚合规则给出以下定理 ：

$$\lambda = GeoMed(\{\tilde{\mathrm{v}}_\mathrm{i}:\ \mathrm{i}\in[\mathrm{n}]\}) \tag{7}$$

定理 2. 令$\{\mathrm{v_i}:\ \mathrm{i}\in[\mathrm{n}]\}$在$R^{\mathrm{d}}$是独立同分布的，$\mathrm{v_i}\sim G$，$E[G]=\mathrm{g}.E\|G-g\|^2=d\delta^2$，

$\{\mathrm{v_i}:\ \mathrm{i}\in[\mathrm{n}]\}$中任意数量的 d 维拜占庭向量$\mathrm{b_1}$，，$\mathrm{b_q}$，如果$\mathrm{q}\le\left[\frac{\mathrm{n}}{2}\right]\text{-}1$，且有

$$\eta_1(\mathrm{n},\ \mathrm{q})\sqrt{\mathrm{d}}\delta<\|\mathrm{g}\|,\ \eta_1(\mathrm{n},\ \mathrm{q})=\frac{2\mathrm{n}-2\mathrm{q}}{\mathrm{n}-2\mathrm{q}}\sqrt{\mathrm{n}-\mathrm{q}} \tag{8}$$

那么，几何中值函数满足$(\alpha_1,\ \mathrm{q})$-维度拜占庭弹性，

其中，

$$\alpha_1\in\left(0,\frac{\pi}{2}\right],\ \sin\alpha_1=\frac{\eta_1(\mathrm{n},\ \mathrm{q})\sqrt{\mathrm{d}}\delta}{\|\mathrm{g}\|} \tag{9}$$

3.4 截尾均值

定理 2 可以证明它满足拜占庭弹性，它在数学期望意义上是收敛的[22].受此启发，本文提出新的聚合规则：截尾均值.

定义截尾均值函数（Trimmed mean）

$$\tau = TriMean(\{\tilde{\mathrm{v}}_\mathrm{i}:\ \mathrm{i}\in[\mathrm{n}]) \tag{10}$$

那么我们给出截尾均值的$(\alpha_2,\ \mathrm{q})$-维度拜占庭弹性定理：

定理 3. 令$\{\mathrm{v_i}:\ \mathrm{i}\in[\mathrm{n}]\}$在$R^{\mathrm{d}}$是独立同分布的，$\mathrm{v_i}\sim G$，$E[G]=\mathrm{g}.E\|G-g\|^2=d\delta^2$，

$\{\mathrm{v_i}:\ \mathrm{i}\in[\mathrm{n}]\}$中任意数量的 d 维拜占庭向量$\mathrm{b_1}$，，$\mathrm{b_q}$，如果$\mathrm{q}\le\left[\frac{\mathrm{n}}{2}\right]\text{-}1$ 且有

$$\eta_2(\mathrm{n},\ \mathrm{q})\sqrt{\mathrm{d}}\delta<\|\mathrm{g}\|,\ \eta_2(\mathrm{n},\ \mathrm{q})=\frac{1}{\mathrm{n}}\left((1-\mathrm{r})(\mathrm{v_{q+1}}+\mathrm{v_{n-q}})+\sum_{\mathrm{i}=\mathrm{q}+2}^{\mathrm{n}-\mathrm{q}-1}\mathrm{v_i}\right) \tag{11}$$

那么几何中值函数满足拜占庭弹性，

其中，

$$\alpha_2\in\left(0,\frac{\pi}{2}\right],\ \sin\alpha_2=\frac{\eta_2(\mathrm{n},\ \mathrm{q})\sqrt{\mathrm{d}}\delta}{\|\mathrm{g}\|} \tag{12}$$

其中，r 代表数据两端截取比例于 n 乘积的小数部分[21].

为了证明定理 3，首先使用下面的引理 2 来限定一维中值.

引理 2 对于由 q 个拜占庭值和 n-q 个正确值组成的序列 v_1， ， $v_{n-q}, q \le \left\lceil \frac{n}{2} \right\rceil - 1$

那么这个序列的均值 m 满足 $m \in \left[\min_i v_i, \max_i v_i\right]$.

条件（1）证明：为了不失去一般性，我假设：

$$E\left[G_i - g_i\right]^2 = \delta_i^2, \quad E\left\|G - g\right\|^2 = E\sum_{i=1}^{d}\left[G_i - g_i\right]^2 = \sum_{i=1}^{d}\delta_i^2 = d\delta^2 \tag{13}$$

$$\left(\tilde{v}_i\right)_j = \begin{cases} \left(v_i\right)_j, & \text{正确节点 j} \\ \text{任意向量}, & \text{拜占庭节点 j} \end{cases} \tag{14}$$

对于 j 维向量，均值：

$$\tau_j \in \left[\min \text{correct}_i\left(\tilde{v}_i\right)_j, \max \text{correct}_i\left(\tilde{v}_i\right)_j\right] \tag{15}$$

因此有：

$$\begin{aligned} & E\left[\tau_j - g_j\right]^2 \le E\left[\max_{correcti}\left(\left(\tilde{v}_i\right)_j - g_j\right)^2\right] \\ & \le E\left[\sum_{correcti}\left(\left(\tilde{v}_i\right)_j - g_j\right)^2\right] = \sum_{correcti} E\left[\left(\left(\tilde{v}_i\right)_j - g_j\right)^2\right] \\ & = \frac{1}{n^2}\left((1-r)(v_{q+1} + v_{n-q}) + \sum_{i=q+2}^{n-q-1} v_i\right)^2 \delta^2 \end{aligned} \tag{16}$$

之后，可以进行如下限定 $\left\|E\left[\tau\right] - g\right\|^2$：

$$\begin{aligned} & \left\|E\left[\tau\right] - g\right\|^2 \le E\left\|\tau - g\right\|^2 \\ & = E\left[\sum_{j=1}^{d}(\tau_j - g_j)^2\right] = \sum_{j=1}^{d} E\left[\left(\tau_j - g_j\right)^2\right] \\ & \le \sum_{j=1}^{d}\frac{1}{n^2}\left((1-r)(v_{q+1} + v_{n-q}) + \sum_{i=q+2}^{n-q-1} v_i\right)^2 \delta_j^2 \\ & = \frac{1}{n^2}\left((1-r)(v_{q+1} + v_{n-q}) + \sum_{i=q+2}^{n-q-1} v_i\right)^2 \sum_{i=1}^{d}\delta_j^2 \\ & = \frac{1}{n^2}\left((1-r)(v_{q+1} + v_{n-q}) + \sum_{i=q+2}^{n-q-1} v_i\right)^2 \delta^2 \end{aligned} \tag{17}$$

之后，可以得到：

$$\langle E[\tau],\ \mathrm{g}\rangle \geq (1-\sin^2\alpha_2)\|\mathrm{g}\|^2 \geq (1-\sin\alpha_2)\|\mathrm{g}\|^2$$
$$\sin\alpha_2 = \eta_2(n,q)\sqrt{d}\delta/\|g\| \tag{18}$$

条件 2 证明：

$$\|\tau\| = \sqrt{\sum_{\mathrm{j}=1}^{\mathrm{d}} \tau_{\mathrm{j}}^{\ 2}} \leq \sqrt{\sum_{\mathrm{j}=1}^{\mathrm{d}} \max_{\text{correct i}} \left(\widetilde{\mathrm{v}}_{\mathrm{i}}\right)_{\mathrm{j}}^{\ 2}} \leq \sqrt{\sum_{\mathrm{j}=1}^{\mathrm{d}} \sum_{\text{correct i}} \left(\widetilde{\mathrm{v}}_{\mathrm{i}}\right)_{\mathrm{j}}^{\ 2}}$$
$$= \sqrt{\sum_{\text{correct1}} \left\|\widetilde{\mathrm{v}}_{\mathrm{i}}\right\|^2} \leq \mathrm{c}_1 \sum_{\text{correct 1}} \left\|\widetilde{\mathrm{v}}_{\mathrm{i}}\right\| \tag{19}$$

之后记下序列 $\{\widetilde{\mathrm{v}}_{\mathrm{i}}:\ \text{correct i}\},\{\mathrm{v}_1,\quad ,v_{n-q}\}.$

那么就可以得出：

$$\left\|\tau\right\|^{\mathrm{r}} \leq \mathrm{c}_2 \sum_{r_1+\quad +r_{n-q}=r} \left\|v_1\right\|^{r1} \quad \left\|v_{n-q}\right\|^{r_{n-q}} \tag{20}$$

这里 C1,C2 是任意的线性组合常数.

综上证明，可以看出截尾均值满足维度拜占庭弹性的两个条件.

3.5 时间复杂度

截尾均值的时间复杂度没有封闭形式的解集[15]，为了计算截尾均值 TriMean（•），我们只需要计算每个维度的中值[16]. 最简单的方法是对每个维度采用任何排序算法[17]，产生时间复杂度 O（dnlogn），它几乎是线性的，且是较低的成本代价.

截尾均值和几何中值同为拜占庭共识中具有强鲁棒性的聚合规则，但是几何中值容易受到极端值影响（如果拜占庭矿工节点提出的向量偏离正确向量很远，那么几何中值就无法较好收敛于一个最优解），而截尾均值在应对极端值时[20]，显得更有鲁棒性，能够收敛于最优解.比如说，假设拜占庭矿工节点知道整个梯度的任何信息，如果这些节点利用梯度总和的大负值缩放[23]，并用缩放后的梯度代替原梯度值，那么几何中值法就无法收敛于最优解，而截尾均值却能够较好收敛于最优解.

4 实验

在本节中我们主要验证所提出的算法的鲁棒性以及收敛性. 我们使用具有两个隐藏层的多层感知器（MLP），在 MNIST 数据集上进行分类，在 n = 30 个工作进程，我们重复每个实验十次，并取平均值. 表 1 列出了数据集的详细信息和相应模型的默认超参数.

4.1 高斯攻击

我们在这个实验中测试经典拜占庭弹性.我们将攻击者用零均值的高斯随机向量和标准差为 200 的协方差矩阵代替一些梯度向量来进行测验的攻击方法，称为高斯攻击.实验结果如表（2）所示，其中的 30 个向量有 9 个是拜占庭向量，所有其他算法之间的差距仍然很小. GeoMed 表现比 TriMean 稍好，其收敛效果较好，而 Krum 收敛速度稍慢，表现最差.

4.2 全方位攻击

我们在这个实验中测试经典拜占庭弹性.假定攻击者知道所有正确的梯度，对于每个拜占庭梯度向量，梯度所有正确梯度由其负和代替.换句话说，这种攻击试图使参数服务器以相反的步长方向进行迭代，其中 30 个梯度向量中有 9 个是拜占庭向量.结果如图 3 所示，Krum 收敛速度较慢，Krum 不如 TriMean 好，但差距很小. 然而，GeoMed 汇聚到了不理想的解集.

表 1.实验概要

Dataset	#train	#test	#rounds	γ	Batch size	Evaluation metric
MNIST	60k	10k	600	0.1	32	Top-1 accuracy

表 2.高斯攻击下不同聚合后的精度

	经过高斯攻击后之后的精度										
聚合规则 迭代轮次	100	150	200	250	300	350	400	450	500	550	600
Krum	0.71	0.82	0.85	0.86	0.87	0.87	0.88	0.89	0.88	0.89	0.89
GeoMed	0.84	0.87	0.88	0.89	0.90	0.90	0.91	0.91	0.92	0.92	0.93
TriMean	0.85	0.87	0.88	0.89	0.89	0.90	0.89	0.90	0.91	0.91	0.92

表 3.全方位攻击下不同聚合后的精度

	经过全方位攻击后之后的精度										
聚合规则 迭代轮次	100	150	200	250	300	350	400	450	500	550	600
Krum	0.80	0.82	0.84	0.86	0.87	0.87	0.88	0.88	0.88	0.89	0.89
GeoMed	0.53	0.63	0.68	0.69	0.69	0.70	0.69	0.70	0.71	0.70	0.71
TriMean	0.81	0.83	0.84	0.85	0.85	0.86	0.88	0.88	0.89	0.90	0.90

4.3 实验小结

我们实验中使用具有两个隐藏层的多层感知器（MLP），在 MNIST 数据集上进行实验的. 我们可以看到，在梯度向量接受高斯攻击情况下，krum 函数并不能保持良好鲁棒性，GeoMed 和 TriMed 可以保持良好的鲁棒性.在接受到全方位攻击下，GeoMed 不能收敛到全局最优解，无法保持鲁棒性，而 TriMed 和 krum 函数却能收敛到最优解，可以保持良好的鲁棒性.

5 总结

本文概括了参数服务器拜占庭弹性，基于随机梯度下降法提出了新的聚合规则—基于均值聚合规则（截尾均值），并证明截尾均值满足维度拜占庭弹性的条件，同时说明它具有接近线性的时间复杂度，最后实验还指出在应对极端值影响时几何中值并不具有鲁棒性，而截尾均值仍然可以收敛于最优解.

致谢

本文受到国家自然科学基金项目（NO. 61672467, 61602418, 61503342）的资助。

References

[1] H.Chen,Y.Li,C.Long,X.Wang Learning Deep Representation for Imbalanced Classification ICCV,2015.

[2] T.Xiao, S.Li,B-C.Wang,L.Lin,X.Wang End-to-End Deep Learning for Person Search ICCV,2015.

[3] Liu Q,Wang L, Huo Q A study on effects of implicit and explicit language model information for DBLSTM-CTC based handwriting recognition[C],IEEE,2015.

[4] Leslie N.Smith,Nicholay Topin Deep Convolutional Neural Network Design Pattern.

[5] E.Ahmed,M.Jones,andT.K.Marks.An improved deep Learning architecture for person re-identification. CVPR,2015.

[6] S. S. Haykin. Neural networks and learning machines, volume 3. Pearson Upper Saddle River, NJ, USA:, 2009.

[7] M. Herlihy, S. Rajsbaum, M. Raynal, and J. Stainer. Computing in the presence of concurrent solo executions. In Latin American Symposium on Theoretical Informatics, pages 214–225. Springer, 2014.

[8] M. Jordan and T. Mitchell. Machine learning: Trends, perspectives, and prospects. Science, 349(6245):255–260, 2015.

[9] Blanchard, Peva, Guerraoui, Rachid, Stainer, Julien, et al.Machine learning with adversaries: Byzantine tolerant gradient descent. In Advances in Neural Information Processing Systems, pp. 118–128, 2017.

[10] Blum, Manuel, Floyd, Robert W, Pratt, Vaughan, Rivest, Ronald L, and Tarjan, Robert E. Time bounds for selection.Journal of computer and system sciences, 7(4):448–461, 1973.

[11] Chen, Tianqi, Li, Mu, Li, Yutian, Lin, Min, Wang, Naiyan, Wang, Minjie, Xiao, Tianjun, Xu, Bing, Zhang, Chiyuan, and Zhang, Zheng. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. CoRR, abs/1512.01274, 2015.

[12] Chen, Yudong, Su, Lili, and Xu, Jiaming. Distributed statistical machine learning in adversarial settings: Byzantine gradient descent. arXiv preprint arXiv:1705.05491, 2017.

[13] Cong Xie ,Oluwasanmi Koyejo , Indranil Gupta.Generalized Byzantinue-tolerant SGD.2018.

[14] H. Mendes and M. Herlihy. Multidimensional approximate agreement in byzantine asynchronous systems. In Proceedings of the forty-fifth annual ACM symposium on Theory of computing, pages 391–400. ACM, 2013.

[15] M. Métivier. Semi-Martingales. Walter de Gruyter, 1983.

[16] D. Newman. Forbes: The World’s Largest Tech Companies Are Making Massive AI

Investments. https://goo.gl/7yQXni, 2017. [Online; accessed 07-February-2017].

[17] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization, 30(4):838–855, 1992.

[18] F. B. Schneider. Implementing fault-tolerant services using the state machine approach: A tutorial. ACM Computing Surveys (CSUR), 22(4):299–319, 1990.

[19] Lynch, Nancy A. Distributed algorithms. Morgan Kaufmann,1996.

[20] McMahan, H. Brendan, Moore, Eider, Ramage, Daniel, Hampson, Seth, and y Arcas, Blaise Aguera.Communication-efficient learning of deep networks from decentralized data. In AISTATS, 2017.

[21] Minsker, Stanislav et al. Geometric median and robust estimation in banach spaces. Bernoulli, 21(4):2308–2335,2015.

[22] Mukkamala, Mahesh Chandra and Hein, Matthias. Variants of rmsprop and adagrad with logarithmic regret bounds. In ICML, 2017.

[23] Seide, Frank and Agarwal, Amit. Cntk: Microsoft's opensource deep-learning toolkit. In KDD, 2016.