# 一种针对毒性攻击的隐私增强联邦学习

## 摘要

联邦学习 (FL) 作为分布式机器学习，近年来受到了相当大的关注。为了缓解隐私问题，FL 本质上承诺多方通过交换梯度而不是原始数据来联合训练模型。然而，FL仍然存在内在的隐私问题，例如，用户的训练样本可以通过仅推断梯度来揭示。此外，新兴的中毒攻击对 FL 也构成了至关重要的安全威胁。特别是，由于 FL 的分布式特性，恶意用户可能会在训练过程中提交精心设计的梯度，以破坏模型的完整性和可用性。此外，同时解决两个问题存在矛盾，即隐私保护 FL 解决方案专门用于确保梯度不可区分性，而针对中毒攻击的防御往往会根据它们的相似性去除异常值。为了解决这样的困境，在本文中，我们的目标是在这两个问题之间建立桥梁。具体来说，我们提出了一个隐私增强 FL (PEFL) 框架，该框架采用同态加密作为基础技术，并通过对数函数的有效梯度数据提取为服务器提供惩罚毒药通道。据我们所知，PEFL是第一个有效检测密文下FL中毒行为的尝试。详细的理论分析说明了该方案的安全性和收敛性。此外，在真实数据集上进行的实验表明，PEFL 可以有效地防御标签翻转和后门攻击，FL 中的两个代表性中毒攻击。索引术语-联邦学习，中毒攻击，隐私保护，云计算。

## 1.概述

联邦学习 (FL) 已成为一种很有前途的机器学习设置，其中多方在本地训练模型的副本，然后将模型更新（即梯度）提交给参数服务器进行聚合，从而构建改进的版本。得益于它在不存储云中的训练数据的情况下训练模型的能力，FL已经成为自动驾驶和医学图像分析等许多安全敏感任务中的基本技术之一。

尽管FL具有潜在的安全和隐私问题，但它也引起了广泛的关注。其中一个突出的问题是梯度的隐私泄露。如以前的研究所示，半诚实的服务器仍然可以通过接收到的梯度恢复目标用户的一些敏感信息(如个人头像、情绪和医疗信息)。FL面临的另一个关键安全威胁是中毒攻击。由于服务器无法访问用户的数据集和联邦训练过程，恶意用户可能会提交定制的梯度从而在测试阶段诱导分类错误。需要注意的是，中毒更新可以控制整个训练过程，从而使最终模型无效。而且这种操作也可能间接侵犯用户的数据隐私，例如，通过上传倒置和放大的梯度，攻击者可以推断样本是否用于训练目标模型。受上述问题的启发，可信的 FL 必须考虑以下基本问题：（1）如何确保用户的数据隐私不被泄露，（2）如何保证模型对对抗性操作的鲁棒性。

为了保护用户的隐私，现有的工作主要集中在确保梯度的机密性。解决方案通常基于以下三个底层技术:差分隐私(DP)、安全多方计算(SMC)和同态加密(HE)。DP是隐私保护分析的数学形式化标准，在效率方面有一些优势的数据使用。然而，基于 DP 的 FL 通常是通过将随机噪声合并到梯度中来完成的，这不可避免地会降低模型的准确性。SMC 通过参与者之间的多个交互实现私有和正确的计算。然而，这种模式不仅会带来巨大的通信成本，而且要求参与者在线保持，这给实际分布式场景带来了新的性能瓶颈。与 DP 和 SMC 相比，HE 允许在训练期间正确聚合梯度，同时支持用户丢失。以前的研究还显示了 HE 实现安全 FL 的卓越性能。

尽管对隐私保护 FL 的主题进行了广泛的研究，但现有的努力并没有同时被认为是一个基本问题，即所有用户是否诚实可靠地提交良性梯度。实际上，目的驱动的恐怖分子可以通过提交恶意精心设计的参数来误导停车标志中的其他自动驾驶汽车向前移动，这对交通安全构成严重威胁。因此，如上所述对于安全且稳健的 FL，在保护用户隐私的同时抵抗恶意对手发起的中毒攻击至关重要。然而，现有的解决 FL 中隐私问题和中毒攻击方案主要集中在两个相反的方向：隐私保护 FL 解决方案试图确保数据不可区分性，而针对中毒攻击的防御倾向于根据它们与良性梯度的相似性去除恶意梯度。目前，相似度主要基于以下一个或多个标准来衡量：

1)欧几里得空间中梯度之间的差异，例如 Krum。

2)梯度分布的变化，如Detox。

3) GeoMed等梯度之间的 L p 范数差异。

从这个角度来看，上述提取有价值统计信息的防御在很大程度上依赖于数据可区分性。矛盾通常使得在不影响用户隐私的情况下识别 FL 中的中毒行为面临重大挑战。我们注意到的与我们的方法最接近的是一篇专注于利用 SMC 使当前基于距离的异常值去除机制与安全聚合兼容的文章。然而，该方案继承了聚合规则的性能和昂贵的计算开销，以及SMC所需的额外通信开销。

为了解决上述挑战，在本文中，我们提出了一个有效且隐私增强的 FL 框架 (PEFL)，它可以在保护隐私的同时有效地检测联邦训练过程中的中毒行为。具体来说，我们工作的贡献有三个方面，如下所示：

（1）我们提出了一种新的隐私增强 FL 框架 (PEFL)，其中 HE 作为底层技术。PEFL 可以通过上传恶意梯度来确保恶意用户无法推断成员资格，此外还可以防止半诚实服务器侵犯用户隐私。

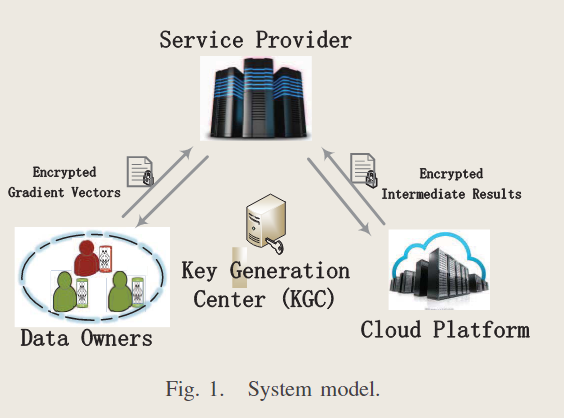
（2）我们还提出了一种新颖的自适应联邦聚合来减轻 FL 中的中毒攻击，它通过以坐标中值作为基准来评估用户的可靠性，然后自适应调整相应用户梯度的权重。为了区分远离基准的恶意梯度和良性梯度，我们进一步提出了一种对数函数来去除恶意梯度。

（3）我们提供了全面的安全分析，并证明了该方案的收敛性。此外，在真实数据集上进行的实验表明，PEFL可以有效地抵抗标签翻转和后门攻击，这两者都是FL中典型的中毒攻击。

本文的其余部分安排如下。在第 2节和第 3节中，我们进行问题概述和初步知识。然后，我们在第 4节中详细介绍了隐私增强联邦学习 (PEFL)，并在第5 节中进行了理论分析。接下来，我们评估了 PEFL 的性能，并分别在第 6 节和第 7 节中讨论相关工作。最后，第 8 节总结了本文。

## 2.问题概述

### A.系统模型如图1所示，我们的系统有四个基本实体：



（1）密钥生成中心 (KGC)：KGC 是独立且受信任的机构，它分发和管理所有公钥和私钥 (pk, sk)。

（2）数据所有者：所有数据所有者，也称为用户，协作训练具有服务提供商协调的统一模型。出于隐私原因，每个用户在设备上的私有数据上本地训练模型，然后将加密的梯度上传到服务提供商。此外，我们假设所有数据所有者持有的数据是独立的同分布的，这与许多先前的工作相同。

（3）服务提供商 (SP)：SP 负责接收用户提交的所有梯度并将它们（通常通过平均）聚合以获得优化的全局模型。同时，需要在云平台的帮助下检测潜在恶意用户发起的中毒攻击。

（4）云平台 (CP)：CP 提供按使用付费的服务。它与 SP 一起工作以执行本文中的计算。此外，CP 持有由可信权威（即本文中的 KGC）生成的私有公钥对 (pk c, sk c)，可用于加密数据或解密密文。

### B.威胁模型

在本文中，我们关注恶意参与者可能滥用的 FL 的隐私和安全漏洞。我们假设每个投毒者都有自己的本地数据集，该数据集来自与其他诚实用户的训练集相同的分布。恶意用户可以操纵存储在其设备上的私有数据以进行数据中毒。与之前的研究类似，为了保证模型的实用性和可用性，我们对恶意用户数量的上限进行了假设，即|F|≤|M|−12，其中F和M分别是恶意用户的数量和总用户，|·|表示集合中用户的数量。此外，由于 SP 和 CP 可以访问所有用户的局部梯度，我们认为它们是半诚实的对手。这意味着 SP 和 CP 在遵守协议执行所有操作时都是诚实的，同时尽可能的根据他们掌握的梯度获得更多信息，从而影响用户的数据隐私。此外，我们假设在之前的研究之后，四个实体(KGC、用户、SP和CP)之间没有共谋。

### C.设计目标

考虑到上述威胁，PEFL 具有以下三个设计目标：

（1）准确性。在本文中，中毒对手提交的参数可能会导致模型对错误分类，从而降低训练模型的准确性。因此，安全和健壮的 FL 需要确保模型的准确性在合理的范围内。

（2）稳健性。鲁棒性要求正确的输出交付给所有协议参与者，无论对手的行为如何。

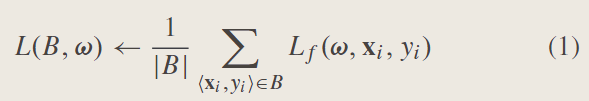
（3）隐私。如先前的研究一样，对手可以通过推断共享梯度来恢复用户的敏感信息，例如训练样本或成员资格。为了保护用户的隐私，必须保密每个用户的局部梯度。

## 3.初步知识

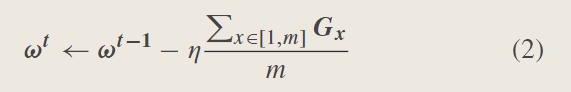
### A.联邦学习

与传统的将所有训练数据集中深度学习不同，FL是一种很有前途的分布式设置，它允许所有数据所有者保持数据本地。在 FL 中，服务器协调训练的整个生命周期，直到模型精度达到所需的水平，或者迭代次数达到预设值。学习的目标是找到最优的模型参数ω\*，使模型的输出y-无限接近给定特征向量x的真实标签。

在本文中，我们专注于监督设置，即用户 U x 持有私有数据集 D x ，其中 D x = { xi , yi ; i =1, 2,…, s}，xi ∈ Rv 表示 v 维特征向量，yi 是对应的类标签。在每次迭代开始时，服务器选择满足资格要求的 m 个用户的子集，例如未占用或连接的用户。每个选定的用户下载当前的模型参数，然后在其私有数据上本地训练模型。为此，我们利用随机梯度下降 (SGD) 算法，这是一种迭代优化学习模型的流行方法。用户不是采用整个数据集，而是在每次迭代中随机选择一个小批量 B 的训练数据来计算损失为：



其中 L f (ω, xi , yi ) 是一个损失函数，用于计算模型当前输出与真实标签之间的差距。然后，计算梯度 G ← ωL(B, ω) 并与服务器共享。最后，服务器聚合所有局部梯度向量并更新全局模型如下：

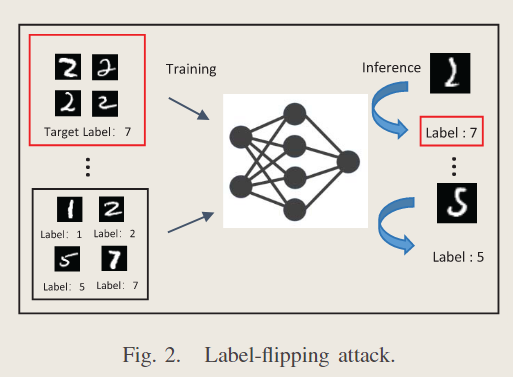


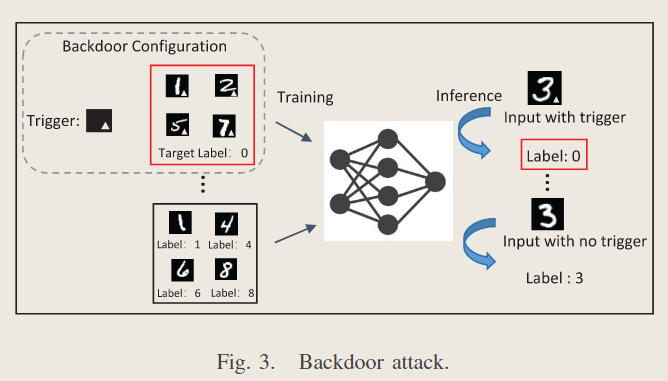
其中ωt表示第t次迭代后的模型参数，η为学习率，{G x, x∈[1,m]}(也在下面的{G x}x=m x=1)为{G1, G2，…, G m}m 个用户梯度的集合。请注意，在本文中，我们使用粗体符号来表示向量，除非有特殊的解释。

### B.毒性攻击

在中毒攻击中，攻击者打算改变训练模型的分类边界，使指定的特征空间映射到目标类。已经证明，单个投毒者可以控制整个训练过程，并损害用户的数据隐私。

在本文中，我们主要关注两种类型的代表性中毒攻击：一种是标签翻转攻击。如图2所示，将正常特征的标签翻转到目标类中。例如，对手错误地将真实标签 2 的样本标记为 7 以进行错误分类。另一个是后门攻击，如图3所示，其目的是寻求一组参数来建立触发器和目标标签之间的强链接，同时最小化对良性输入分类的影响。例如，对手期望这样一个任何以帽子为触发器的图片都将被归类为攻击者期望的类别，而没有帽子的图片将被正确分类的模型。为此，Bagdasaryan等人提出了一种优化模型，通过将项目添加到局部模型的损失函数中来最小化与原始模型参数的距离:αl后门+(1−α)l，其中l个后门用于引入后门，l是原始参数和新参数之间的均方误差。





### C.线性同态加密

同态加密允许直接在加密数据上执行计算，而不需要访问密钥。在本文中，我们利用线性同态加密(LHE)来实现隐私增强FL，这是一种公钥加密方案，支持对密文进行线性同态运算。一般来说，LHE 由算法元组 (KeyGen, Enc, Dec, Eval) 组成，如下所示：

（1）HE.KeyGen(1k) → (pk, sk)。HE。KeyGen表示为概率多项式时间(PPT)算法，其输入为安全参数1k，输出为公钥pk和私钥sk。

（2）HE.Enc(pk, x) → c。在接收到公钥pk和明文x后，加密函数HE.Enc 输出对应于 x 的密文 c。为简洁起见，本文采用[[x]]pk表示用公钥pk加密x的密文形式。

（3）HE.Dec(sk, c) → x。HE.Dec 是一个解密函数，它将私钥 sk 和密文 c 作为输入，并返回对应于 c 的明文 x。

（4）HE.Eval(pk, c1, c2, f L) → c。给定公钥pk和密文c1、c2以及线性函数f L，评估函数HE.Eval 输出密文，使得 HE.Dec(sk, c ) = f L (x1, x2)，其中 ci = Enc( pk, xi ) 对于 i ∈ {1, 2}。

LHE满足语义安全和功能隐私。语义安全意味着给定两个从 {0, 1}l 和密文 c 中采样的消息 x1 和 x2，多项式时间对手 A 将密文 c 与 x1 或 x2 区分开来，概率最多可以忽略不计超过二分之一。功能隐私要求同态操作在其功能上没有透露不必要的信息。例如，我们采用Paillier密码系统，它是一种具有代表性的加性同态加密，具有以下两个基本同态性质，即给定两个明文x1和x2、常数r和公钥pk，我们有:

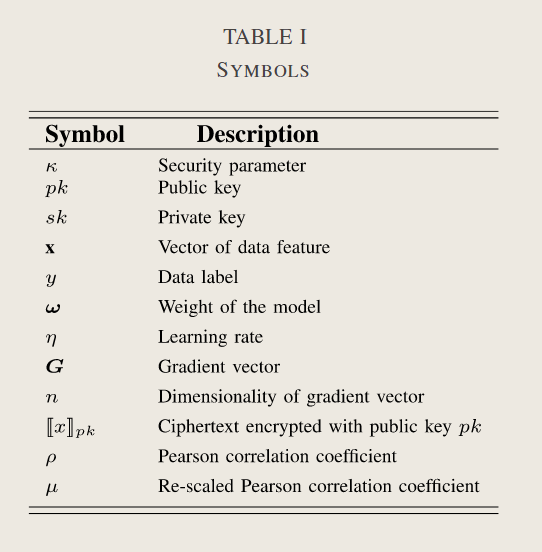
(a) [[x1]]pk·[[x2]]pk = [[x1 + x2]]pk;

(b) [[x1]]r pk = [[r·x1]]pk

我们强调优化 LHE 是一个超出本文范围的正交方向，我们提出的方案可以从 LHE 的任何性能改进中获得直接收益。

## 4.解决方案

符号表示：



### A.概述

联邦学习的核心在于假设用户持有的局部训练数据是独立的同分布的(独立同分布简称为i.i.d.)，这允许局部梯度是全局更新的无偏估计。该假设也是 PEFL 的基础，它试图识别和阻止偏离诚实多数梯度的梯度向量。

我们的深入见解是，恶意的用户提交恶意梯度，以寻找一组具有恶意目标的参数ω τ，这与诚实用户的目标模型ω\*不同，即ω\* ≠ ω τ。这意味着恶意梯度向量和良性梯度向量之间存在可察觉的差异。实际上，与良性梯度的低相似度意味着梯度是高概率恶意的，这是许多相关研究的共识，我们识别和阻断恶意梯度。在本文中，我们通过计算梯度之间的 Pearson 相关系数来衡量相似性，这是高维变量之间最常用的相似性度量之一。公式如下：



我们可以从上面的公式中观察到 Pearson 相关系数等于协方差 Cov(X, Y )，除以两个变量 σ (X)σ (Y ) 的标准差的乘积。

为了提高效率，我们建立了一个基准（即本文中的坐标中值）来区分异常行为和诚实多数。然后，我们设计了一个对数函数来提取有效信息，并自适应地赋予每个梯度的权重。特别是，相关性较低的梯度被认为是异常的，而它们的权重应该设置为零，以获得模型的准确性和鲁棒性。为了保护用户的隐私，我们进一步提出了四种安全协议来聚合梯度向量和优化密文域中的共享模型，这是通过采用Paillier密码系统的加性同态特性来实现的。

### B.构建PEFL

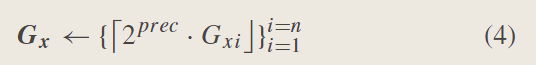
1)系统设置:我们需要一个可信密钥生成中心(KGC)来为云平台(CP)生成LHE的一对非对称密钥(pk c, sk c)，其中私钥k c仅由CP保存。同时，所有授权用户都持有KGC生成的LHE的一对非对称密钥(pk x, sk x)。此外，在协议开始时，服务提供商 (SP) 随机初始化全局模型 ωini t 的参数。

2)安全训练:安全训练过程包括两个阶段:局部训练阶段和鲁棒聚合阶段。详细步骤描述如下：

a) 局部训练阶段：在这个阶段，我们假设子集中毒用户比例的上限与全部中毒用户的比例上限相同，即小于 50%。在第 t 轮，对于所有 x ∈ [1, m]，用户 U t x 接收具有公钥 pk x 的加密模型 [[ωt ]]pk x。在解密它们后，U tx 训练并获得局部梯度向量 G tx。

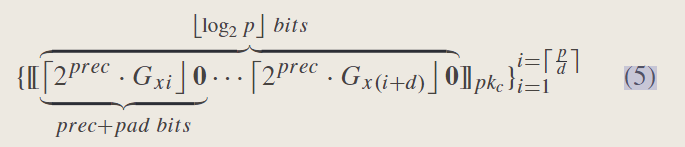
我们没有单独上传当前梯度，而是使用动量的 SGD 来平滑更新，它使用指数衰减因子 γ (0 < γ < 1) 添加了一系列先前的梯度。因此，用户提交的参数变为：G x = ∑ l∈[0,t ] γ t−l G lx。El-Mhamdi等人表明，动量不仅可以加速收敛，还可以降低方差范数比，有利于模型的鲁棒性。PEFL 在减轻 FL 中的中毒攻击方面继承了这一优势。

为了保护数据隐私，U t x 使用 CP 的公钥 pk c 加密梯度向量 [[G x ]]pkc。由于加密原语的基本要求，for i ∈[1,n]，梯度向量的每个条目G xi应编码为整数形式，如下所示:



其中 [a] ∈ Z 表示到实数 a 最接近的整数，prec 是精度的位。此外，G xi 是梯度向量 G x 中的第 i 个条目。

为了减少通信成本，我们利用将多个明文打包成一个密文的密文打包技术，并使用单指令多数据(SIMD)技术并行对这些值进行操作。假设公钥pk c = p(一个大的正整数)，对于明文空间log2 p≥2048位，每个用户将d个明文打包成一个密文，如下所示:

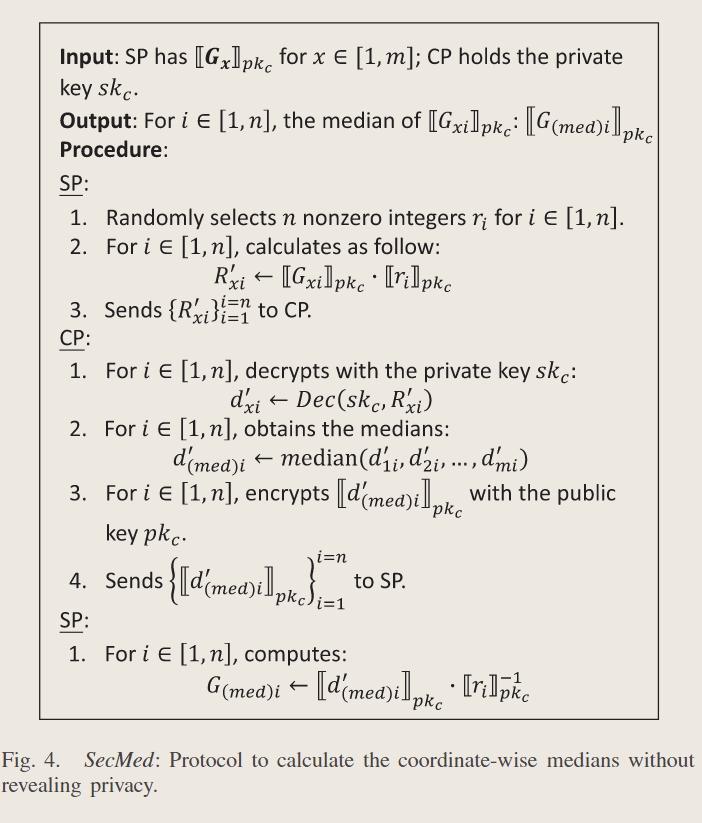


其中 0 表示填充位的零填充，以防止在密文添加中溢出。此外，d =⌊log2 p/(prec+ pad)⌋，取实数的整数部分，⌈p/d⌉ =⌊p/d⌋+1。为了简单起见，我们使用以下[G xi表示2prec·G xi]0。U t x最终将打包密文 [[G x ]]pkc = {[[G xi · · · G x(i+d)]]pkc }i=p/d i=1 发送给 SP。

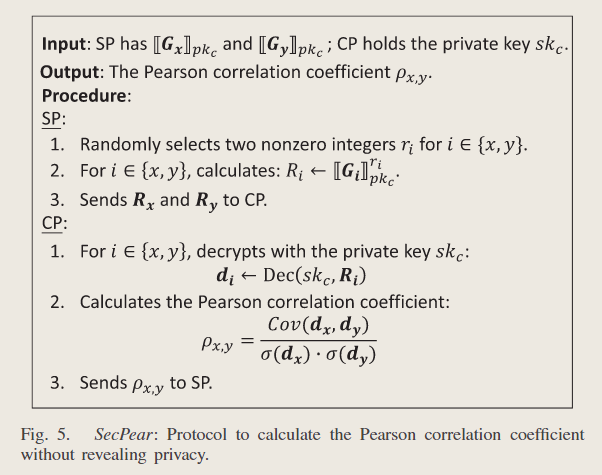
b)鲁棒聚合阶段:在这个阶段，SP与CP交互，识别和阻止恶意用户发起的中毒攻击。值得注意的是，我们是第一个在密文下实现鲁棒聚合的人。

首先，在接收到来自m个选定用户的加密梯度向量{[[G x]]pkc}x=m x=1时，SP与CP一起工作，计算坐标中值{[[G(med)i]]pkc}i=n i=1，而不透露{G x}x=m x=1。为了实现这一点，我们设计了安全协议 SecMed。

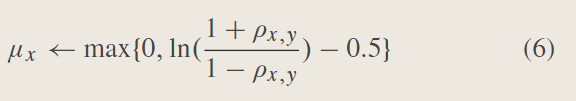
SecMed 的具体步骤如下：对于 x ∈ [1, m]，SP 首先通过将随机采样值 ri 的密文相乘来模糊 {[[G xi ]]pkc }i=n i=1，并将模糊值 Rxi 发送给 CP。然后 CP 计算每个坐标的中值，即对于 n 维向量，需要计算 n 个中值。接下来，CP 对中位数进行加密，并将它们发送给 SP。最后，SP 消除噪声以获得所需的结果。图 4 显示了协议 SecMed 的详细信息。



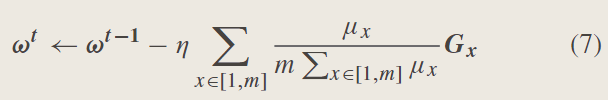
在 PEFL 中，坐标中值被视为基准，即我们计算坐标中值与用户 U t x 的梯度之间的 Pearson 相关系数 ρx,y。为了在不泄露隐私的情况下计算相关性，即梯度向量仍然对 SP 和 CP 保密，我们提出了一种安全的协议SecPear。具体来说，SP首先通过计算梯度向量的密文到第r次幂来模糊梯度向量，其中r是随机采样的非零整数。在解密上述密文后，CP计算模糊变量的相关性，并将结果发送给SP。图5描述了SecPear协议。正确性和安全性分别在第 4-C2 和 5-A 节中讨论。



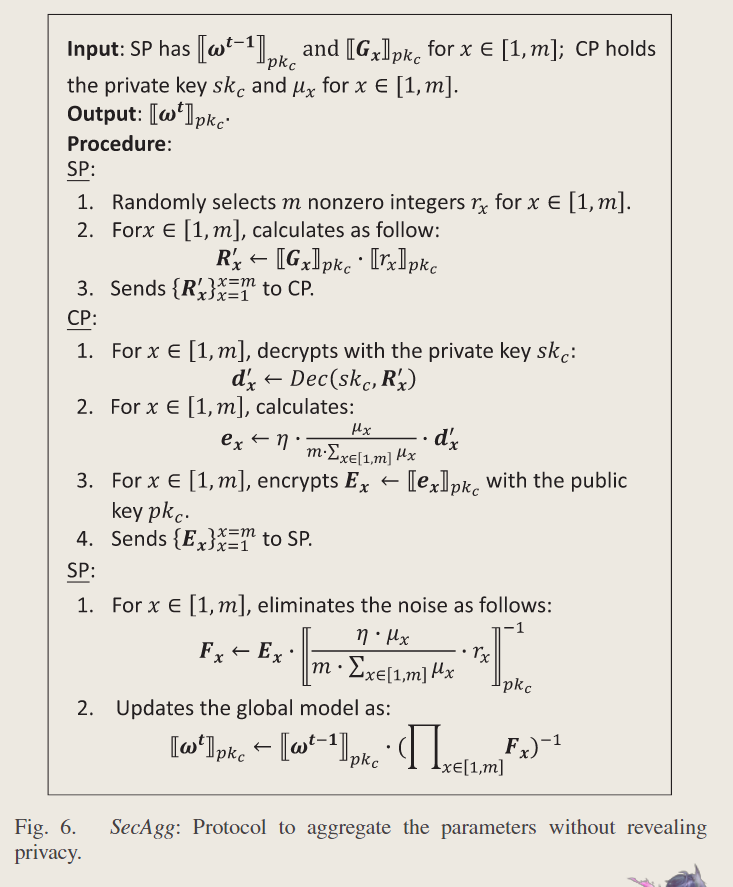
对于 m 个局部梯度，SP 调用 m 次协议SecPear 以获得 m 个相关系数。然后，每个系数重新缩放如下：

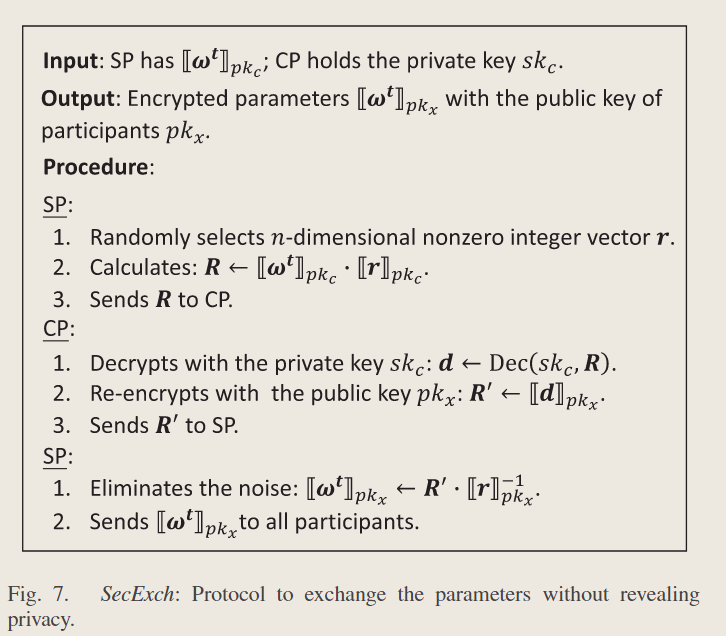


回想一下，我们认为相关性较低的用户是毒药者。公式(6)中的对数函数进一步期望函数的两个尾部附近的值具有更高的散度，从而阻碍恶意行为。同时，尽可能惩罚相关性较低的诚实用户。给定重新缩放的值，每个梯度向量都分配权重。然后 SP 在密文域下更新共享模型，如下所示：



协议SecAgg 旨在实现上述过程，而不会泄露 G x。图 6 说明了细节。最后，SP 与 CP 通信，以获得重新加密的全局参数 [[ωt ]]pk x 和所有用户持有的公钥pk x，然后广播给所有用户进行进一步训练。图 7 显示了协议 SecExch 的详细信息。





备注：在本文中，我们认为 Pearson 相关系数的隐私不需要受到保护。很明显，即使系数是公开的，服务器也无法反转梯度的大小和符号。最近，Geiping等人提出梯度与梯度值的相似度包含了更多关于训练数据的信息。他们还设计了一种基于梯度之间的余弦相似度的优化方案来重建训练样本。然而，该方案在很大程度上依赖于梯度的符号，这是 PEFL 中的隐藏信息。因此，我们的方案使数据重构攻击无效。我们承认系数也可以用来进行一些攻击，例如模型窃取攻击。但是，它超出了本文的范围。这个问题仍然开放。

### C.讨论

在这里，我们讨论如何优化上述协议，并详细证明协议SecPear.

1)优化的正确性:我们使用以下优化来减少基本协议的通信开销。

-删除冗余：观察到 m 中存在冗余，需要协议 SecPear 来计算用户提交的梯度向量与坐标中值之间的 Pearson 相关系数。SP在协议SecPear的每个调用上重复向CP发送中值。因此，在优化的协议中，SP 仅为重复项发送一次。

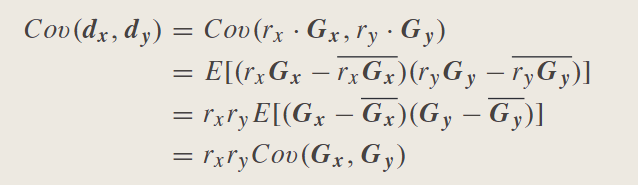
-减少通信轮次：我们观察到协议SecPear 和 SecAgg 中存在两个独立的关系。一种是对于所有 x ∈ [1, m],R x 在协议 SecPear 中是相互独立的。另一个是协议 SecPear 中的 R x 和协议 SecAgg 中的 R x1 是相互独立的。因此，对于 x ∈ [1, m]，SP 通过单轮发送 R x 和 R x1 来减少通信轮次。

2)正确性：为了确保 PEFL 能够有效地识别恶意梯度，我们需要确保仍然可以从协议 SecPear 中的模糊数据中正确获得梯度之间的相关系数。

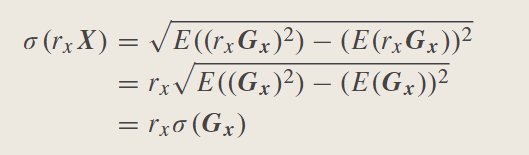
命题 1（正确性）：给定扰动梯度向量，协议 SecPear 可以正确获得用户提交的梯度向量与其坐标中值之间的 Pearson 相关系数。

证明：根据加密原语的同态，我们有：SecPear 中的 dx = rx·G x。为了证明正确性，它等价于证明：ρx,y = ρdx ,dy。

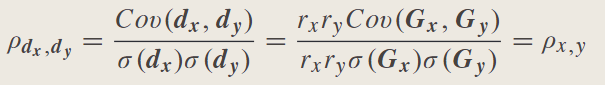
根据第 4-A 节中 Pearson 相关系数的定义，我们分别考虑变量的协方差和标准偏差。首先，协方差公式如下：



其中 E(G x ) 表示变量 G x 的期望，G x- 是变量 G x 的平均值。类似地，标准偏差随随机变量的大小成比例，如下所示：



因此，方程满足如下：



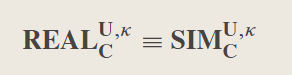
总之，即使给定模糊变量，也可以正确计算 Pearson 相关系数。

## 5.分析

### A.安全属性

LHE的功能隐私和语义安全(简称IND-CPA安全性)保证了协议的强安全性，有效的防御了选择明文攻击。在本节中，我们提供了一个依赖于模拟器的混合参数，以进一步证明在协议执行期间，除了从计算结果推断的信息外，服务器的联合视图不会泄露用户私有数据的任何信息。

命题 2（诚实但好奇的安全，带有好奇的服务器）：给定一个安全参数 κ，用户 U 和 C = {S P, C P} 的任何子集，令 REALU,κC 是一个随机变量，表示 C 中各方在上述协议的实际执行中的联合视图。存在一个概率多项式时间 (PPT) 模拟器 SIM，使得 SIM 的输出在计算上与 REALU,κC 无法区分：



证明：根据 REALU,κC 的定义，它由 C 中各方在执行协议期间收到的所有内部状态和消息组成。我们采用中使用的标准混合参数来证明这个命题，即给定安全参数κ，我们通过对REALU、κC中的随机变量的一系列(多项式地多)后续修改来定义PPT模拟器SIM，使得SIM的输出在计算上与REALU、κC无法区分。详细的证明如下所述。

Hyb1 ：我们初始化一个随机变量，其分布与真实协议执行中 C 中各方的联合视图 REALU,κC 无法区分。

Hyb2：在这种混合中，我们改变了模拟诚实用户 U x ∈ U 的行为，以便每个用户 U x 使用 LHE 的公钥 pk c 加密随机选择的向量 βx，而不是原始梯度向量 G x。由于只改变了密文的内容，LHE的 IND-CPA 安全属性以及两个非串通 SP 和 CP 设置保证了这种混合与之前的混合无法区分。

Hyb3：在这种混合中，我们模拟了SP通过随机均匀采样的噪声ζi来扰动βxi，而不是[[Gxi]]pkc·[[ri]]pkc。众所周知，均匀随机数添加的参数也是均匀随机的。由于噪声ζ的分布是均匀随机的，因此前一个混合参数也是均匀随机的，所以这里的这种混合和前一个混合是从相同的分布中采样的，即均匀随机的。此外，LHE 的 IND-CPA 安全属性以及两个非串通 SP 和 CP 设置保证了这种混合与之前的混合无法区分。

Hyb4：在这种混合中，我们将SP和CP执行的协议SecPear的输入更改为[βx]]pkc和[[θi]]pkc而不是[[Gx]]pkc和[[G(med)i]]pkc。LHE 的 IND-CPA 安全属性，以及两个非串通 SP 和 CP 设置保证了这种混合与之前的混合无法区分。

Hyb5：在这个混合中，对于所有用户 U x ∈ U ，我们模拟 SP 来计算 [[βx ]]ζx pkc 而不是 [[G x ]]r x pkc ，其中 x ∈ {1, m}。虽然 CP 持有可用于解密上述密文的私钥 sk c，但乘以随机数的参数仍然均匀随机，这与之前的混合一致。因此，这种混合与前一个混合没有区别。

Hyb6：在这种混合中，我们将SP和CP执行的协议SecAgg的输入更改为[ξt−1]]pkc和[[βx]]pkc，而不是[[ωt−1]]pkc和[[Gx]]pkc。由于只改变了密文的内容，LHE 的 IND-CPA 安全性以及两个非串通 SP 和 CP 设置保证了这种混合与之前的混合无法区分。

Hyb7：这种混合类似于Hyb3，对于所有用户U x∈U，我们模拟SP通过随机噪声ζx而不是[[G x]]pkc·[[rx]]pkc来扰动βx。LHE 的 IND-CPA 安全属性，以及两个非串通 SP 和 CP 设置保证这种混合与前一个混合分布相同。因此，这种混合与前一个混合没有区别。

Hyb8：在这种混合中，我们将SP和CP执行的协议SecExch的输入更改为[ξt]]pkc，而不是[[ωt]]pkc。LHE 的 IND-CPA 安全属性以及三个非串通实体（即 SP、CP 和用户）设置保证了这种混合与之前的混合无法区分。

Hyb9：与Hyb3类似，我们不是发送：[r+ωt]]pkc=[[ωt]]pkc[[r]]pkc，而是替换发送到CP的参数，以使用[ξt]]pkc[[ζt]]pkc重新加密。即使持有私钥sk c的CP也可以解密上述密文，混合和(r + ωt)具有相同的分布，即均匀随机。因此，这种混合与前一个混合没有区别。

论证证明有一个从上述分布中采样的模拟器 SIM，以便其输出在计算上与 REAL 的输出无法区分。因此，PEFL 持有好奇的 SP 和 CP 对用户的私有数据没有任何了解的安全属性。

命题 3：PEFL 具有恶意用户无法破坏其他用户隐私的安全属性。

证明：在我们的威胁模型中，恶意用户作为主动对手发起攻击并损害其他用户的隐私。如之前的研究所示，他们通过利用错误的梯度向量来推断训练数据集中是否存在目标样本。攻击平均在很大程度上取决于聚合规则。在本文中，每个梯度向量都是自适应给定的权重，而不是平等地对待它们。根据第6-B2节的实验结果，PEFL将攻击的成功率降低到0.04，即恶意用户无法推断在训练过程中是否使用目标样本。因此，PEFL 具有恶意用户无法破坏其他用户隐私的安全属性。

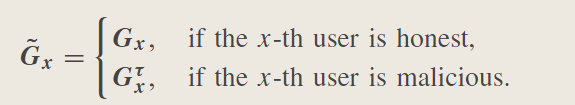
### B.收敛性

现在，我们证明所提出方案的收敛性。

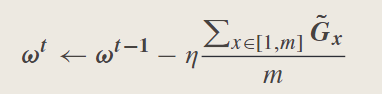
要求4(误差项):恶意梯度和良性梯度之间存在误差项e，使∑x∈F Gτx =∑x∈H G x +e成立。

证明:我们推广了中毒攻击的过程，寻找一组具有恶意目标的模型参数ω τ，这与诚实用户的目标模型ω\*不同，即ω\*≠ ω τ。设ω\*和ω τ之间的距离为h-，即h- = ω\*−ω τ。请注意，恶意用户和诚实用户只能通过提交梯度来实现他们的目标。

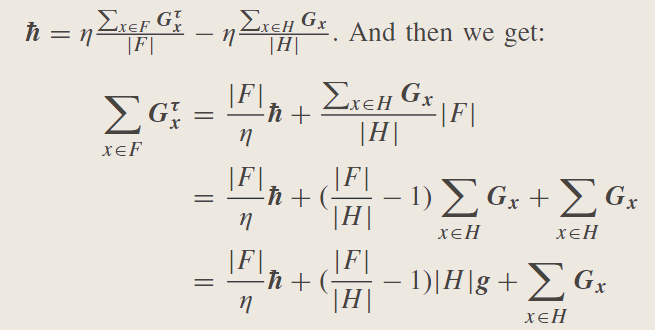
现在我们正式定义经典的中毒攻击模型。假设在特定回合中，从随机变量 G ← ωL(B, ω) 中抽取的正确梯度向量 {G x , x ∈ [1, m]} 是独立同分布的 (i.i.d.) 样本，其中 E[G] = g 是梯度的无偏估计量。因此，对于任何 x ∈ [1, m]，E [G x ] = E[G] = g。对于恶意用户，SP接收到的实际向量{̃G x, x∈[1,m]}如下:



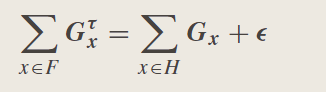
请注意，SGD 算法的收敛特性的一个重要假设是每个良性梯度都是实际梯度的无偏估计，这通常是通过均匀随机抽样来保证的。然而，恶意用户提交的梯度打破了均匀随机抽样的限制。因此，恶意梯度向量和良性梯度向量之间存在明显的差异。根据用于更新模型参数的 SGD 算法的定义，我们有：



给定恶意用户的梯度 {Gτx , x ∈ F}，其中 F 表示恶意用户的集合，恶意用户的目标是获得以下形式的参数：ω τ ←ωt -1 − η∑x∈F Gτx|F |。类似地，给定诚实用户的梯度 {G x , x ∈ H }，其中 H 表示诚实用户的集合，诚实用户旨在获得以下形式的参数：ω\* ← ωt -1 − η∑x∈H G x|H |。根据之前定义的参数距离（即h- = ω\* - ω τ ），我们有：



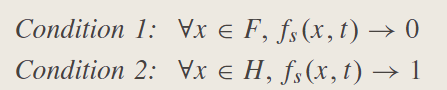
因此，由于目的不同，恶意梯度和良性梯度之间存在误差项，即



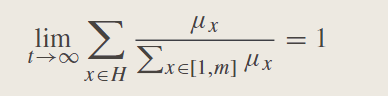
其中e = |F |/η \*h- + ( |F |/|H | − 1)|H |g。

命题 5（收敛）：给定第 4-B 节中描述的 PEFL 的详细步骤，恶意或/和诚实用户的收敛速度在 T 次迭代上为 O( 1/T^2 )。证明：PEFL 遵循自适应学习率方法的轮廓，该方法已应用于 SGD 算法并提供收敛保证。众所周知，具有恒定学习率的 SGD 算法的收敛速度为 O( 1/T^2)。

假设用户U x的自适应学习率ηx = η × μx∑x∈[1,m] μx由函数fs (x, t)决定，其中t表示当前迭代。U x 来自 F 或 H ，分别表示协议中恶意用户和诚实用户的集合。如前所述，本文恶意用户的百分比小于 50%，即 2|F| < |F| + |H|。SGD的收敛特性可以应用于我们的解决方案，只要用诚实的用户数据进行训练，满足以下条件:



我们认为 G x 是初始模型 ωini t 相对于 U x 本地数据优化模型 ω x 的理想梯度。根据要求 4，我们知道： ∑x∈F Gτx =∑x∈HG x +。在这方面，PEFL可以根据差异识别异常，降低恶意梯度的权重。如果实现了恶意目标，随着迭代次数的增加而增加。这导致重新缩放的系数接近 0，即 fs (x, t) → 0。因此，中毒者的权重为 0，满足条件 1。因此，PEFL 对诚实用户的影响满足：



也就是说，诚实用户的权重之和接近1，满足条件2。上述证明表明PEFL满足这两个条件。因此，PEFL的收敛速度与具有自适应学习率的SGD相同，即O(1/T^2)。

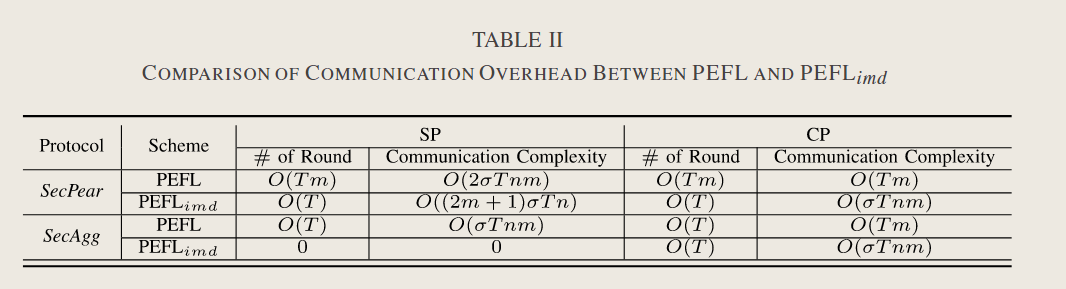
### C.效率评估

我们现在讨论 PEFL 的通信和计算开销。如第 4-C1 节所述，我们优化了基本 PEFL 的通信开销。在下文中，我们将 PEFLimd 表示为优化方案。

1)通信开销:在PEFL中，SP与CP交互，执行四个安全协议进行安全聚合。需要提到的是，使用密文打包技术的通信开销增加因子在σ = 2(1 + pad/prec)左右，如之前研究中所讨论的。例如，我们取pad = 15和pr ec = 32，因此通信因子增加2.93。在SecMed中，SP与CP相互作用得到n个坐标中值，其中n表示梯度向量的维数。因此，SP 和 CP 的通信开销分别为 O(σ T nm) 和 O(σ T n)，其中 T 是迭代次数，m 表示用户数。对于每个用户，协议 SecPear 在每次迭代中调用一次，因此通信轮次为 O(T m)。我们注意到 SP 在 SecPear 中向 CP 发送两个加密向量，这导致 O(2σ T nm) 的通信成本。对于 CP，它将明文中的 m Pearson 相关系数发送给 SP。因此，通信轮次和成本分别为 O(T m) 和 O(T m)。然后，调用协议SecAgg安全地更新模型参数。

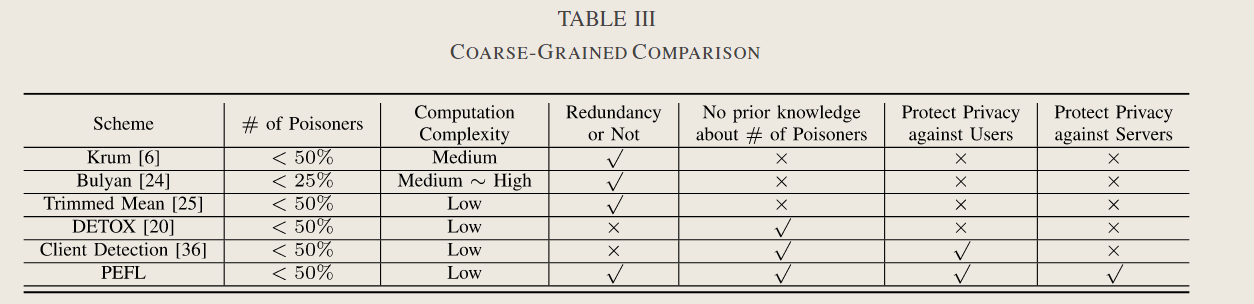
对于每次迭代，SP 和 CP 都以密文的形式相互通信 m 个模糊梯度向量，因此通信轮次和成本相同，O(T ) 和 O(σ T nm)。最后，对于 SP 和 CP，协议 SecExchin 分别遵循相同的通信轮次和成本，即 O(T ) 和 O(σ T n)。

在 PEFLimd 中，SP 和 CP 仅交互一次以通信协议 SecPear 和 SecAgg 中使用的所有数据。因此，协议 SecPear 中 SP 和 CP 的通信轮次应该是 O(T)，SP 的通信轮次称为协议 SecAgg。在每一轮通信中，SP 以两种方式发送已掩盖的加密梯度向量，以及它们的坐标中值。因此，SP 的通信开销为 O((2m + 1)σ T n)。CP 发送的数据与 PEFL 中的数据相同。所以通信开销是相同的。如表 II 所示，我们可以观察到 PEFLimd 在通信开销方面优于基本的 PEFL。



2)计算开销:在PEFL中，主要的局限性在于鲁棒聚合阶段的计算成本。在这个阶段，SP 与 CP 交互以计算密文域中 m 个梯度向量的坐标中值、Pearson 相关系数和权重。为了计算坐标中值，最简单的方法是每个维度运用一些复杂度为 O(nm log m) 的排序算法。在 PEFL 中，我们利用 BFPRT 算法获得中位数，这是一种著名的算法来解决从 m 个数字中选择第 k 个最大或第 k 个最小数的经典问题，最坏情况的时间复杂度为 O(m)。因此，协议 SecMed 的计算成本平均需要线性时间 O(nm)。对于每个用户，我们只需要一次计算相关性。因此，SecPear协议产生的计算开销也是O(nm)。此外，PEFL 的参数更新阶段不会引入额外的计算复杂度，除了调整需要 O(m) 计算复杂度的权重。总体而言，防御阶段产生的计算复杂度为 O(T nm)。在用户端，计算成本随着模型参数的数量线性增加。这使得每个用户的计算成本为 O(T ns)，其中 s 是每个用户持有的样本数量。

与 Krum 和 Bulyan相比，计算成本随着用户数量的增加呈二次增长，PEFL 在计算成本方面更有利。我们还在计算开销方面与其他五个最先进的作品进行了粗粒度比较，如表 III 所示。



### D.功能性

在这里，我们通过与 Krum [6]、Bulyan [24]、Trimmed Mean [25]、Detox [20] 和客户端检测 [36] 五种最先进的鲁棒方案进行比较来分析 PEFL 的功能优势。表 III 显示了毒药数量、计算复杂度、冗余与否、毒药数量的先验知识以及隐私方面的粗粒度比较。

由于 Bulyan 是 Krum 的改进版本，Bulyan 对投毒者数量的限制更严格，计算复杂度更高。Krum 和 Bulyan 每次迭代的运行时间通常在用户数量上是二次的，甚至更高。虽然 Trimmed Mean 和 PEFL 在计算上是有效的，但在用户数量上几乎是线性的。Detox 和 Client 检测的计算复杂度也与用户的数量呈线性关系，因为它们都需要大量冗余来过滤几乎所有的恶意梯度。

此外，Krum、Bulyan 和 Trimmed Mean 的设计要求中毒者的数量作为先验知识，有助于提高防御的有效性。在实际场景中，这种假设是不现实的，即对手的身份和数量无法提前感知。对于Detox和Client检测中的每个中毒子模型，超过一半的用户需要报告异常。所以他们不需要这些信息作为先验知识。在 PEFL 中，我们专注于设计一个实用且有效的鲁棒聚合规则，因此我们的方案没有考虑合理的假设。

客户端检测的解决方案声称保护用户的数据隐私免受恶意用户的攻击，这是通过使用差异隐私的可信服务器实现的，以扰乱模型参数不被泄露。然而，差分隐私机制和客户端检测都会导致误报概率的增加。此外，服务器并不总是可信的。根据之前的研究，服务器可以破坏用户的数据隐私。在PEFL中，LHE和两个非串通服务器设置的IND-CPA安全特性保证了用户的数据隐私。安全证明在第 5-A 节中给出。

## 6.性能评估

### A.实验准备

我们从两个方面展示了 PEFL 的性能：

a.验证有关准确性的声明。

b.验证PEFL对真实攻击的鲁棒性。

为了突出 PEFL 的优势，除了三种典型的鲁棒方案，即 Krum [6]、Bulyan [24]、Trimmed Mean [25]，我们建立了一个对照组——未修改的 FL 作为基线，取平均值作为聚合规则。

1)攻击:我们主要关注两种具有代表性的中毒攻击:标签翻转攻击和后门攻击。为了模拟标签翻转攻击，我们将恶意用户持有的源类重新标记为目标类。在我们的实验中，源类和目标类分别为 1 和 9。为了重现后门攻击，在每一轮训练中，恶意用户从私有数据集中随机采样一定数量的图像（例如 1000），并以最大强度覆盖右下角的 5 × 5 像素。所有这些修改后的图片都将它们的标签重置为目标类，例如 CIFAR-10 数据集的飞机。5 × 5 模式被视为触发器。

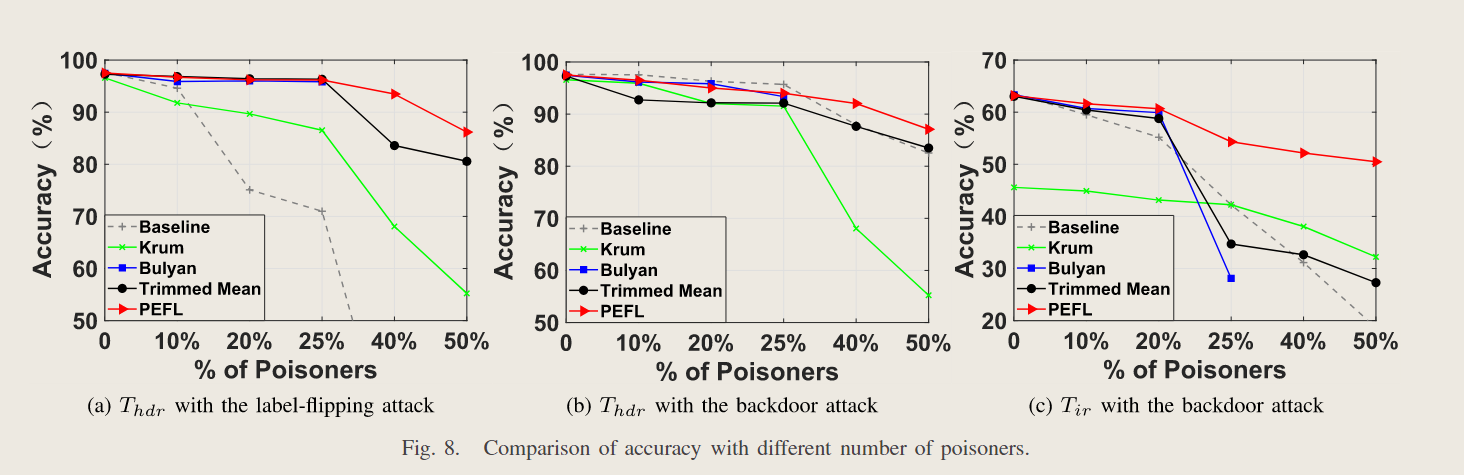
2)数据集和模型架构:为了评估PEFL的性能，我们模拟了两种场景:手写数字识别(Thdr)和图像识别(Tir)。Thdr 旨在根据学习到的特征将手写数字分类为 10 个类别（0-9）之一。在这个任务中，我们使用一个简单的两层全连接网络作为模型结构，每一层的参数设置依次为724×100和100×10。该数据集是经典的 MNIST 数据集，包括 60,000 张大小为 28 × 28 的手写数字图片和 10,000 个测试样本。对于 Tir，我们的实验是在 CIFAR-10 数据集上实现的，该数据集共有 60,000 张彩色图像。这些图像的大小是 32 × 32，分为 10 个类别（例如飞机、汽车和鸟类），每个类别有 6000 张图像。有 50,000 片用于训练，另外 10,000 片用于测试。该任务的模型架构由两层卷积和三层全连接组成。此外，通过执行 torch.utils.data.distributed.DirtributedSampler() 将训练数据分发给每个用户。

3)超参数:我们在实验中使用的用户总数为51。对于每个实验，我们设置批量大小为2 7，动量为0.9，初始学习率为0.1。每个报告的数据平均需要5个实验。

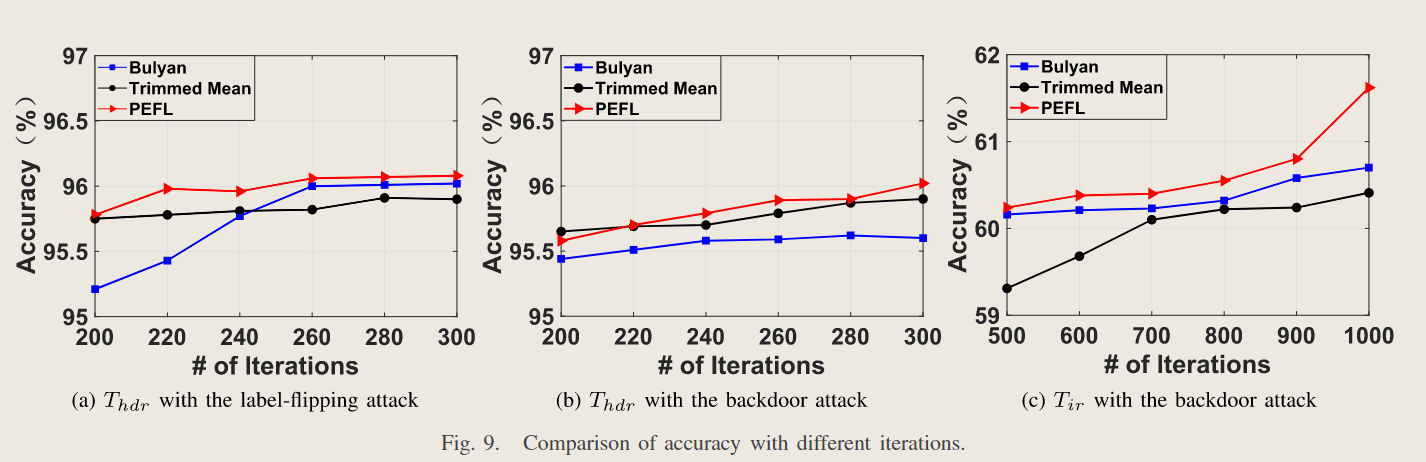
### B.实验结果

1）准确性评估：许多因素会影响模型的准确性，包括模型容量、数据质量和数据量、迭代次数和恶意用户的比例。在我们的实验中，我们使用公共数据集在两个应用场景中评估我们的解决方案的性能，因此我们只通过滑动后两个因素的值来展示模型的准确性，冻结模型的容量并忽略数据质量和数量的问题。值得一提的是，Tir 的模型架构相对简单，而具有更复杂特征的 CIFAR-10 数据集欠拟合，导致模型精度低。然而，我们在这里只关注不同防御之间的准确性差异。在下文中，我们描述了这两个因素对模型精度的影响。

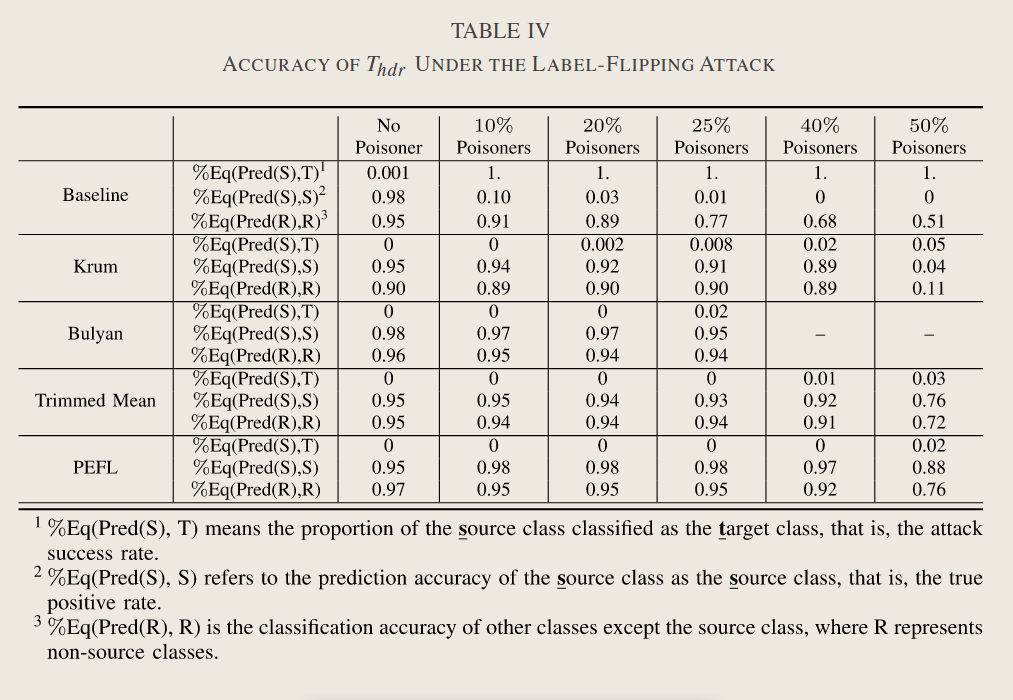
a) 不同比例的投毒者的影响：直观地说，训练中诚实用户持有的更良性的数据带来了更好的模型精度。先前的工作表明，即使只有一个恶意参与者，具有平均聚合规则的 vanilla 联邦学习也很容易受到攻击。图 8 概述了测试准确度与不同数量的投毒者的比较。图中的 Bulyan 是不完整的，因为投毒者比例方案的上限为 25%。我们观察到，随着投毒者数量的增加，所有防御的准确性都会降低。我们分析了这是因为随着投毒者比例的增加，训练过程中有价值的数据量减少，导致模型精度下降。在中毒比例超过 25% 后，我们可以发现下降程度变大。然而，无论应用场景如何，与其他防御相比，PEFL 仍然可以保持比较大的优势。我们将其归因于坐标中值和围绕中值聚合参数的好处。另一个有趣的事情是 PEFL、Bulyan 和 Trimmed Mean 的准确率明显高于 Krum。一个合理的解释是聚合参数的数量。如前所述，Krum 只选择一组候选参数进行更新，丢弃剩余的 (m-1) 梯度向量集。虽然其他人尽可能地聚合良性参数。

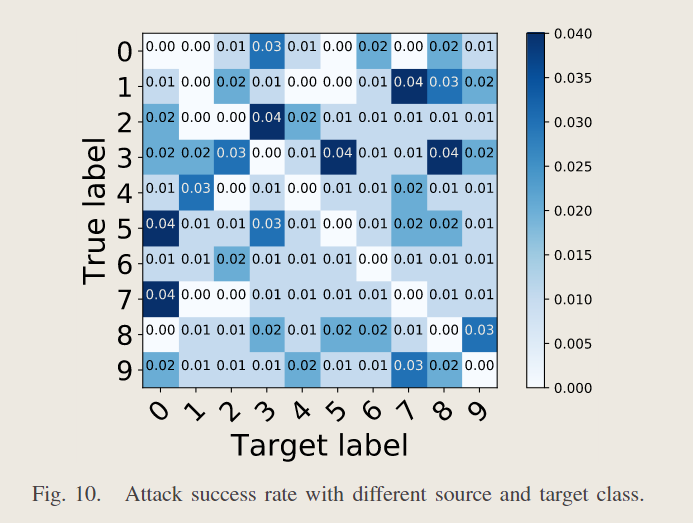


b) 不同迭代的影响：随着迭代次数的增加，模型可以提取更有效的数据特征，这也会导致模型更准确的测试精度。在这里，我们忽略了模型容量过小或大导致的欠拟合和过拟合的问题。图 9 显示了精度与不同迭代的比较。此外，由于 Bulyan 对投毒者比例的限制，图 9a 和图 9b 固定了 25% 的投毒者，即51个用户中总共混合了 12 个投毒者。图9c固定了20%的投毒者，如图8c所示，当投毒者的比例为20%时，三种方案的准确率非常接近。我们可以观察到 PEFL 的准确率略高于 Bulyan 和 Trimmed Mean。这是因为 PEFL 调整每个梯度的权重，进一步提高了剩余参数的可靠性，有利于模型的收敛性。

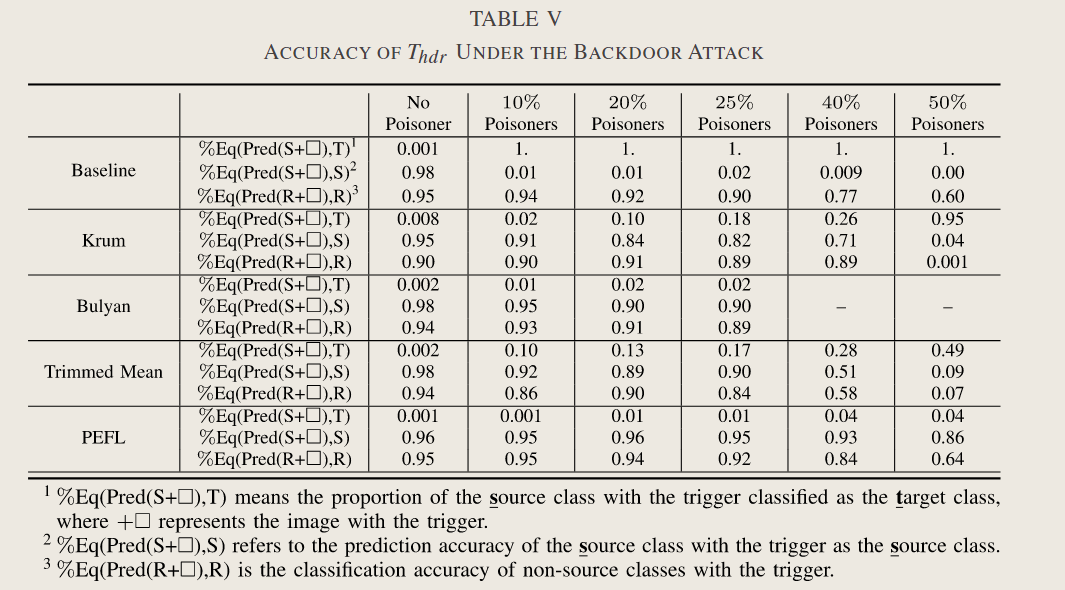


2)鲁棒性评价:本文的鲁棒性被认为是防御中毒攻击的能力。我们通过描述攻击的成功率来评估防御方案的鲁棒性。对于标签翻转攻击，我们通过观察测试集中识别的目标类的比例来测试攻击的成功率。在表 IV 中，我们展示了对手的攻击成功率、源类和非源类在不同毒药比例下的分类准确性，其中非源类是指除源类之外的所有类的集合。我们观察到，即使只有少数毒药师，基线也非常脆弱，攻击成功率高达 100%。这也突出了我们的研究方向的重要性。观察实验结果，我们得到即使恶意用户持有的源类被翻转并标记为目标类，源类的分类精度略高于其他非源类。我们之所以这样分析，是因为标签为 1 的源类的数据特征比实验中的其他数据更容易识别。为了验证PEFL对不同源类和目标类对的防御效果，图10列举了不同源类和目标类的攻击成功率(即误分类为目标类的源类的比例)，当毒药者的概率达到50%时，其中y轴表示源类，x轴表示对手指定的目标类。我们可以观察到最大攻击成功率为 0.04，这也表明 PEFL 在标签翻转攻击下是稳健的。



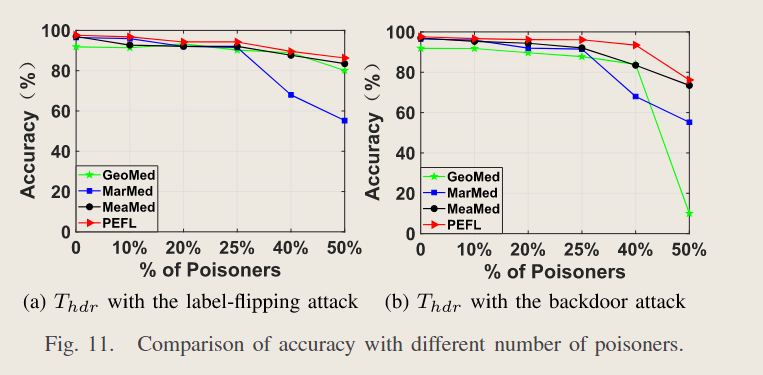


与标签翻转攻击不同，对于后门攻击，我们将触发器添加到测试样本中，并观察带有触发器的图片分类。表 V 列出了每种方案在不同比例的恶意用户下三组数据：以触发器分类为目标类的目标类、分类为目标类的触发器的图片比例以及正确分类带有触发器的非目标图片的比例。我们的实验结果表明，随着中毒比例的增加，PEFL 仍然可以有效地防御后门攻击并保持高精度。



总之，根据实验结果，我们可以清楚地看到，在标签翻转攻击和后门攻击下，PEFL 在防御能力和分类准确性方面优于其他解决方案（即 Krum、Bulyan 和 Trimmed Mean）。

3)与基于中位数的方案的比较:PEFL的有效性主要取决于坐标中值的可靠性。因此，我们提供了与三种基于中值的方案的比较，即 GeoMed、MarMed、MeaMed。根据前人研究中的定义，GeoMed 表示所有梯度向量的地理中值，MarMed 将坐标中值作为更新，MeaMed 平均最接近坐标中值的前 (m − f ) 值，其中 f 是用户子集中的恶意用户。图11描述了两种攻击方法下不同毒药剂比例下四种方案的Thdr的准确性。由于 Tir 具有相似的结果，我们在这里仅使用 Thdr 来展示 PEFL 的性能优势。我们分析了 GeoMed 和 MarMed 精度较低的原因主要是每次迭代只将一个用户的梯度作为更新。与 Krum 类似，它忽略了剩余 (m − f − 1) 个诚实用户的贡献。当协议中多达 50% 的毒药剂时，GeoMed 不会收敛。与 GeoMed 和 MarMed 相比，MeaMed 在这两种攻击下具有更高的预测精度。然而，MeaMed 的设计需要毒药剂的数量作为先验知识。如前所述，这是一个不切实际的假设。



与三种基于中值的方案相比，PEFL 在攻击下都具有最好的准确度优势，并且不需要提前知道毒药剂的数量。总之，PEFL 优于上述基于中值的方案。

## 7.相关工作

### A.鲁棒性聚合规则

针对联邦学习中的中毒攻击，机器学习社区提出了几种防御方案。一般来说，现有的解决方案主要涉及以下三种方法。从数据级别来看，基于统计的聚合允许参数服务器根据从训练数据或梯度中提取的统计数据聚合参数，而不是对它们进行平均。代表性示例包括基于欧几里得距离的 Krum [6]，以及基于地理中值的 GeoMed [21]。然而，这些方法对于训练精度和效率付出相当大的代价。从模型级别，参数服务器将更新后的子模型分配给数据所有者，以确定是否存在异常[20]、[36]。显然，这种方法不可避免地会产生冗余的通信开销。更糟糕的是，它只能容忍有限数量的毒药剂。最近，一些研究人员提出从神经元层面断开从特定特征空间到目标标签空间的端到端映射。例如，王等 提出了 Neural Cleanse，它通过发现连续改变模型输出的小输入扰动来识别木马行为的存在，然后通过模型修补算法修剪神经元。然而，这种方法需要神经网络专家和大量的干净样本。

由于与本文的相关性，我们在这里提出了更多基于统计的防御，其关键思想是尽可能地将聚合规则集中在良性梯度向量上。例如，Blanchard 设计了 Krum [6]，它选择与其他梯度最相似的梯度之一作为全局更新。这里的相似性是通过计算最接近梯度向量的 (m − f − 2) 梯度的欧几里得距离之和来衡量的。Krum 选择总和最小的梯度。Bulyan [24] 本质上是一个结合了 Krum 和 Trimmed 均值的变体。Bulyanfirst 选择与 Krum 相同的规则小于 (m − 2 f ) 梯度，然后将最接近梯度向量中值的参数平均为全局更新。然而，Fang等人[40]认为，上述两种聚合规则对于具有一定知识的对手都不够有效，他们可以精心设计一组相似的梯度来混淆聚合规则。Trimmed mean [25] 删除了一些极值，然后将剩余的参数平均为全局更新。GeoMed [21] 将梯度的中值作为全局更新。然而，它们的计算复杂性随着梯度的维度呈指数增长。此外，Fang等人[40]提出了一种基于用户参数的均值和方差来伪造梯度的最大值和最小值，将上述聚合规则拉到陷阱中。

然而，上述基于统计的防御会损害模型的准确性。在本文中，我们提出了具有令人满意的精度的 PEFL，主要受益于坐标中值。

### B.隐私加强的联邦学习

最近，启用隐私保护FL主要基于以下三个底层技术:差分隐私(DP)[12]、安全多方计算(SMC)[14]、同态加密(HE)[18]和可信执行环境(TEE)[41]、[42]。然而，上述每种方法都有其局限性，这仍然是隐私保护 FL 的实施是一个悬而未决的问题。

具体来说，Shokri 和 Shmatikov [12] 首先提出了一种隐私保护 FL 框架，该框架通过选择性地共享模型的一小部分参数并通过利用 DP 机制扰动它们来实现。然而，该方案必须在准确性和隐私之间进行权衡。为了进一步提高模型的效用，Abadi等人[10]设计了矩会计来跟踪DP训练过程中隐私损失矩的边界。然而，Jayaraman 和 Evans [19] 表明，当前基于 DP 的工作很少提供可接受的效用-隐私权衡。Payman [26] 和 Bell 等人通过使用秘密共享来保护训练数据的隐私，这是实现 SMC 的代表性技术。但是，它需要用户在线停留，这给实际分布场景带来了新的性能瓶颈。最近，Tramer和Boneh[42]在TEE中提出了神经网络的私有执行，通过使用专门的硬件将敏感计算与不受信任的软件隔离开来。不幸的是，这种方法很难大规模扩展，因为它的硬件成本昂贵且可扩展性稀缺。

上述方案仅限于 FL 中的隐私问题。在本文中，我们在隐私保护和防御中毒攻击之间建立了可行的桥梁，为未来FL框架的设计提供了新的思路。

## 8.结论

在本文中，我们提出了一个名为 PEFL 的新框架，它可以在保证对代表性数据中毒攻击的鲁棒性的同时提供隐私保护联邦学习。通过综合理论分析证明了PEFL是安全的和收敛的。此外，实验结果表明，PEFL 在准确性和鲁棒性方面具有可比性。在未来的工作中，我们将专注于提高 PEFL 的准确性，并进一步利用优化效率的方法。