RCP:本地差分隐私下的均值保护技术

刘利康, 周春来

引用本文

刘利康, 周春来. RCP:本地差分隐私下的均值保护技术[J]. 计算机科学, 2023, 50(2): 333-345.

LIU Likang, ZHOU Chunlai. RCP:Mean Value Protection Technology Under Local Differential Privacy [J]. Computer Science, 2023, 50(2): 333-345.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700273

RCP:本地差分隐私下的均值保护技术

刘利康　周春来
中国人民大学信息学院　北京 100872
(micahliu2012@gmail.com)

摘　要　文中主要围绕差分隐私查询中的均值估计问题展开论述,介绍了目前主流的数值型数据均值估计的本地差分隐私设计方案,首次引入随机响应技术中的随机截尾机制来揭示本地差分隐私下均值计算的基本原理,提出了关于均值估计方差的效用优化定理,给出了边界优化公式,从而提高了该领域效用优化理论的可解释性和可操作性。基于该理论,首次提出了一种实用、简洁、高效的均值估计算法协议 RCP,可用于收集和分析连接到互联网的智能设备用户的数据,同时满足本地差分隐私要求。RCP 构造简单,支持在任意数量的数值属性上执行数据分析任务,通信与计算高效,有效缓解了现有算法设计复杂、优化困难、效率较低等实际问题。最后,通过实证研究证明了所提方法在效用、效率和渐进误差界限上优于现有的其他方案。

关键词:本地差分隐私;均值估计;随机响应;随机截尾;效用优化
中图法分类号　TP309

RCP:Mean Value Protection Technology Under Local Differential Privacy

LIU Likang and ZHOU Chunlai
School of Information,Renmin University of China,Beijing 100872,China

Abstract　This paper mainly focuses on the mean estimation problem in differential privacy query. After introducing the current mainstream local differential privacy design scheme of numerical data mean estimation,it first introduces the random censoring mechanism in random response technology to reveal the basic principle of mean calculation under local differential privacy,proposes a utility optimization theorem about the variance of mean estimation,and gives a boundary optimization formula,which improves the interpretability and operability of utility optimization theory in this field. Based on this theory,this paper proposes a practical,concise and efficient mean estimation algorithm protocol RCP for the first time,which can be used to collect and analyze the data of intelligent device users connected to the Internet,while meeting the requirements of local differential privacy. RCP is simple in structure,supports data analysis tasks on any number of numerical attributes,and has efficient communication and calculation,effectively alleviating the practical problems of complex algorithm design,difficult optimization,and low efficiency. Finally,empirical research demonstrates that the proposed method outperforms other existing schemes in terms of utility,efficiency and asymptotic error bounds.

Keywords　Local differential privacy,Mean estimation,Random response,Random censoring,Utility optimization

1　引言

差分隐私(Differential Privacy,DP)是由 Dwork 等[1-4]提出的,并在多个领域已成为保护隐私的事实标准。传统的差分隐私也称为中心差分隐私(Centralized Differential Privacy,CDP),它首先收集用户的原始数据,然后将扰动后的聚合信息发布给公共用户。它假设数据管理员是可信的,但现实世界中并不总是如此,即使是声誉良好的大公司也无法保证其客户的隐私。

为解决这一问题,本文提出了本地差分隐私(Local Differential Privacy,LDP)[5]。图 1 给出了中心差分隐私(见图 1(a))和本地差分隐私(见图 1(b))框架的比较。对于中心差分隐私,数据管理员(Curator)拥有用户的真实数据;而在本地差分隐私模型下,管理员持有扰动后的数据而非原始数据,查询是在扰动后的数据集上执行的。因此,本地差分隐私可以防止不受信任的数据管理员泄露隐私,减轻受信任的数据管理员保持数据安全的负担。

如图 1 所示,目前对于主流差分隐私框架的计算,本文主要关注本地框架下数值型数据扰动与均值估计方法,并在以下几个方面做出了贡献。

(1)对随机响应技术中的随机截尾模型(Random Censoring Model,RCM)进行了详细分析,揭示了当前主要的 LDP

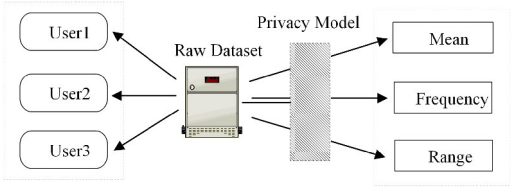
到稿日期:2022-07-27　返修日期:2022-10-21
基金项目:国家自然科学基金(61732006)
This work was supported by the National Natural Science Foundation of China(61732006).
通信作者:周春来(czhou@ruc.edu.cn)

均值计算方法的基本原理，并提出通过非对称尺度区间调节来获取估计方差最优界限的效用优化定理和优化公式。

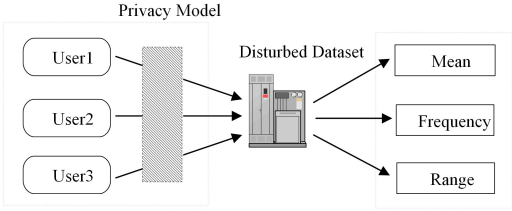
(2)利用上述效用优化理论设计出比现有同类型算法效用更高、计算与通信成本更低、系统构成更为简洁、可操作性更强的本地差分隐私均值估计技术 RCP。

(3)实现了基于单属性(维)数据的 RCP 算法协议(RCP_Single)和多属性(维)数据的 RCP 算法协议(RCP_Multiple)，并通过重构尺度区间约束方程分别实现了它们的效用优化版本。

(4)在有关智能设备的真实数据集上进行实验，并采用 Friedman 秩和分析法对所提算法的性能进行严谨的定量分析，在实证研究中证实了 RCP 优化算法与其他 4 种同类型的算法之间在性能上存在显著性差异；并且在不同隐私预算的性能指标检验中，RCP 优化算法的平均排名始终为第一，从而得出 RCP 的性能优于对比算法的结论。



(a)中心差分隐私



(b)本地差分隐私

图 1 两种差分隐私模型的比较

Fig. 1 Comparison of two differential privacy models

2 相关定义

2.1 差分隐私与本地差分隐私

根据文献[1-2]的描述，差分隐私中的定义用于保证在给定原始数据集上看到输出的可能性接近于在任何一行记录上与原始数据集不同的另一数据集上看到相同输出的可能性。

定义 1(ε-差分隐私) 随机算法 $\mathcal{M}$ 满足 ε-差分隐私(ε-DP)，当且仅当对于在一个元素上不同的任何相邻数据集 D 和 D' 以及任何可能的输出 $y\in\Omega$，有：

$$\Pr[\mathcal{M}(D)=y]\leqslant e^{\varepsilon}\Pr[\mathcal{M}(D')=y] \tag{1}$$

其中，ε 是分配给所讨论差分隐私机制的隐私损失预算，ε 的取值范围为 $[0,+\infty)$，ε 越小，隐私信息的不可区分性越高，隐私保护越好。

定义 2(ε-本地差分隐私) 随机算法 $\mathcal{M}$ 满足 ε-本地差分隐私(ε-LDP)，当且仅当对于任何一对输入值 $x,x'\in D$ 和任何可能的输出 $y\in\Omega$，有：

$$\Pr[\mathcal{M}(x)=y]\leqslant e^{\varepsilon}\Pr[\mathcal{M}(x')=y] \tag{2}$$

本地环境下并无受信任第三方。每个用户在将自己的数据发送到服务器端之前都会对其进行扰动，因此即使服务器端是恶意的，且与所有其他参与者共谋。根据 LDP 的保证，个人的隐私数据仍然会受到保护。更多的 LDP 资讯可参阅文献[6-7]。

到目前为止，LDP 的扰动机制基本源自于随机响应的概念，这是一种在社会科学中已有数十年历史的技术，有关概念可以参考文献[8]。

2.2 本地差分隐私下的均值估计

首先考虑用户只有一个数值型属性项 A 的情况，其取值用 $x\in\mathbb{R}$ 表示。

为获取 LDP 下的均值，一个朴素的想法是使用拉普拉斯机制。为方便计算，将用户的原始值 x 归一化到 $[-1,1]$，然后将拉普拉斯噪声 $\mathrm{Lap}(2/\varepsilon)$ 添加到该值，由分布的特性可知估计无偏，很容易计算出该估计量产生的预期误差为 $O(1/(\varepsilon\sqrt{n}))$。但在多个属性项的情况下，若设属性数量为 m，则预期误差为 $O(m/(\varepsilon\sqrt{n}))$，显然当 m 很大时，效用会急剧下降。

Duchi 等从理论上研究了 ε-LDP，提出了一种极值扰动(Extreme Values Perturbation，EVP)的随机化机制，该机制已扩展到许多场景[9-11]。该方案主要为：用户根据输入值 x 以指定的概率报告两个极值中的一个，无论 x 是什么，该概率确保估计值 $\hat{x}$ 的期望等于真值 x，因此最后得到 x 均值的无偏估计。当 ε 很小时，该方法的方差极小，但当 ε 变大时则效果不好。Kairouz 等证明了随着 ε 的增加，极值扰动机制并非总是最优的[12]。

Wang 等提出了一种分布扰动机制，称为分段机制(Piecewise Mechanism，PM)[13]。具体来说，定义一个比输入域 $[-1,1]$ 更宽的输出域 $[-s,s]$。对于每个值 x，都有一个相关的范围 $[l(x),r(x)]$，其中 $-s\leqslant l(x)\leqslant r(x)\leqslant s$。用户以高概率输出 $\hat{x}\in[l(x),r(x)]$，以低概率输出其他值。与 Duchi 的方法相比，当 ε 较大时方差会小很多。因此，PM 机制在低隐私态下获得了更好的精度，缺点是由于输出的取值范围无界，因此计算复杂并且难以编码。

Li 等提出了一种与 PM 类似的方法，称为方波机制(Square Wave Mechanism，SW)，用于重建分布而不是直接计算均值[14]。给定输入域 $[0,1]$ 和输出域 $[-b,1+b]$，用户以概率 $e^{\varepsilon}/(2be^{\varepsilon}+1)$ 输出 $\hat{x}\in[x-b,x+b]$，以概率 $1/(2be^{\varepsilon}+1)$ 输出其他值。通过最大化报告输入和输出之间的互信息上限来选择参数 b。由于输入值总是处于高概率区域的中心，因此无法提供均值的无偏估计。

此外，Nguyen 等在极值扰动的基础上提出了 Harmony 方法，在多属性状态下随机选择一个属性进行报告，而非报告所有属性值[15]。

Harmony 构建理念与 Duchi 的方法类似，隐私保证和渐近误差界限与 Duchi 的方法相同，但通信成本低得多。

3 本地差分隐私下的随机截尾机制

极值扰动方法本质上仍属于经典随机响应的一环，其中随机截尾机制能很好地诠释此类思想的内涵。

3.1 随机截尾模型

随机截尾模型（Random Censoring Model，RCM）[16-18]与Warner模型[8,19-20]类似，也是一种调查敏感性问题属性特征的方法，不同的是随机截尾模型用于调查定量特征而非定性特征。

假设用户数量为 n，每个用户只有一个数值型属性项 $x\in[a,a+b](b>0)$ 且 x 取值的分布未知，因此用随机变量 X 表示 x。

通常情况下计算样本均值 $\frac{1}{n}\sum_{i=1}^{n}x_i$ 即可估计期望 EX 的大小。但由于随机算法 $\mathcal{M}$ 作用于 x，因此我们仅能获取扰动值 $\mathcal{M}(x)$。于是采用如下方法估计 EX。

设 X 的概率密度为 $\varphi(x)$，并设 Y 为服从 $[a,a+b]$ 上均匀分布的随机化器，Y 的概率密度为 $\varphi(y)=1/b$，对每个 x_i 的扰动方式如下：

使用随机化器 Y 生成值 $y_i\in[a,a+b]$，如果 $x_i\geqslant y_i$，则报告1，否则报告0。将 x_i 扰动后的值用 Kronecker 记号表示：$1_{x_i\geqslant y_i}$，于是：

$$E(1_{X\geqslant Y})=\Pr(X\geqslant Y)=\int_a^{a+b}\int_y^{a+b}\varphi(x)\phi(y)\mathrm{d}x\mathrm{d}y$$

$$=\frac{1}{b}\int_a^{a+b}(x-a)\varphi(x)\mathrm{d}x=\frac{EX-a}{b}\tag{3}$$

易知 $1_{x_i\geqslant y_i}$ 服从两点分布，设为 Z_i。同时设

$$Z=\frac{1}{n}\sum_{i=1}^{n}Z_i=\frac{1}{n}\sum_{i=1}^{n}1_{x_i\geqslant y_i}\tag{4}$$

$$EX=\mu,\frac{1}{n}\sum_{i=1}^{n}x_i=\hat{\mu}\tag{5}$$

由于 $\hat{\mu}$ 是 μ 的无偏估计，因此 $E(bZ+a)=bE(1_{X\geqslant Y})+a=EX$。最后得到：

$$\hat{\mu}=bZ+a\tag{6}$$

由期望 $EZ_i=\frac{x_i-a}{b}$ 和方差 $Var(Z_i)=\frac{(x_i-a)(a+b-x_i)}{b^2}$ 得到 $\hat{\mu}$ 的估计方差：

$$Var(\hat{\mu})=Var(bZ+a)\leqslant\frac{(\hat{\mu}-a)(a+b-\hat{\mu})}{n}\tag{7}$$

3.2 构建满足差分隐私的随机截尾机制

不妨设 $a=0,b=1$，则 $\hat{\mu}=Z$。这时 $x_i\in[0,1]$，对于任意 x_i，Y 随机均匀抽取 $y_i\in[0,1]$，y_i 有 x_i 的概率落在 $[0,x_i]$，有 $1-x_i$ 的概率落在 $(x_i,1]$，即：

$$\Pr(Z_i=k|Y=y_i)=x_i^k(1-x_i)^{1-k}$$

其中，$k=0,1$，考察 $\forall y_1,y_2(y_1\neq y_2)$ 的概率比：

$$\frac{\Pr(Z_1=1|Y=y_1)}{\Pr(Z_2=1|Y=y_2)}=\frac{x_1}{x_2}\tag{8}$$

固定 $x_1\in(0,1]$，令 $x_2\to0$，则：

$$\Pr(Z_2=1|Y=y_2)\to0$$

因此，式(8)$\to+\infty$，故经典随机截尾模型并不满足 ε-差分隐私。为了构造具有差分隐私性的扰动机制，我们对概率范围进行适度放缩，建立 $[0,1]$ 到 $[c,c+d]$ 的线性映射，其中 $0<c<c+d<1$。令 $W=dX+c$，此时 Y 与 W 的联合分布如图2所示。

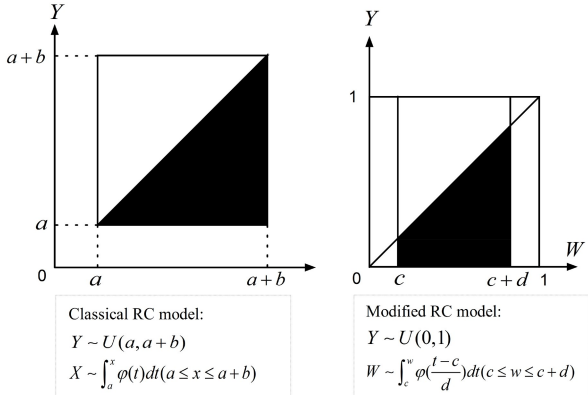


图2 经典随机截尾模型与修改后的随机截尾模型的比较

Fig. 2 Comparison between classical random censoring model and modified random censoring model

图2中左图为修改前的RC模型，y 处于黑色区域时 $\mathcal{M}(x)=1$，y 处于白色区域时 $\mathcal{M}(x)=0$；图2中右图为修改后的RC模型，$w\in[c,c+d]$，同样 y 处于黑色区域时 $\mathcal{M}(x)=1$，y 处于白色区域时 $\mathcal{M}(x)=0$。

用户抽取的 y_i 有 $w_i=dx_i+c$ 的概率落到 $[0,w_i]$，有 $1-w_i$ 的概率落到 $(w_i,1]$，于是：

$$\hat{\mu}=\frac{Z-c}{d}\tag{9}$$

$\hat{\mu}$ 的估计方差为：

$$Var(\hat{\mu})\leqslant\frac{(d\hat{\mu}+c)[1-(d\hat{\mu}+c)]}{d^2n}\tag{10}$$

由 $\min(dx+c)=c$，$\max(dx+c)=d+c$，再次考察概率比函数：

$$\frac{\Pr(Z_1=1|Y=y_1)}{\Pr(Z_2=1|Y=y_2)}=\frac{dx_1+c}{dx_2+c}\leqslant\frac{d+c}{c}=e^{\varepsilon}$$

得到关系式：

$$R(c,d,\varepsilon)=ce^{\varepsilon}-c-d=0\tag{11}$$

该式建立了线性映射的尺度区间 $[c,c+d]$ 与隐私预算 ε 之间的关系，我们称之为满足 ε-差分隐私的随机截尾机制的约束方程。

因此，我们利用 $R(c,d,\varepsilon)=0$ 可以创建出满足 ε-差分隐私的随机截尾估计模型。显然，在本地环境中，基于该模型构造的随机算法 $\mathcal{M}$ 一定满足 ε-LDP，我们将其称为 ε-LDP下的随机截尾隐私机制（Random Censoring Privacy Mechanism，RCPM），简记为RCP。

4 高效的均值估计算法协议 RCP

在我们的应用环境中，服务器端从一个数量为 n 的智能设备用户收集数据，并计算所收集数据的统计模型。我们的目标是最大限度地提高模型的准确性，同时保护用户的个人隐私。

一个自然的前提是服务器已经知道终端的身份，如IP地址，但没有它们的隐私数据。收集的过程并不会产生更多信息增量来改变这个事实，而差分隐私算法协议保证了这一点。

形式上，假设每个用户 $u_i(i=1,\cdots,n)$ 只发送一个数据样本，u_i 的隐私数据由属性列 $(A_1,\cdots,A_m)$ 表示，m 被称为属性的维数，属性的取值记为 $(x_{i1},\cdots,x_{im})$。

不失一般性，我们假设每个数据元素 $x_{ij}(j=1,\cdots,m)$ 都

来自实数数域 $\mathcal{D}=[0,1]$，$\mathcal{D}$ 也被称为“数据字典”。

当 $m=1$ 时，用户数据只有一个属性，简记为 A，对应取值记为 x_i。我们将此时用户的属性状态称为单属性态；相应地，当 $m\geqslant 2$ 时称为多属性态。

本文提供的算法协议框架主要包括“用户端”和“服务器端”两部分，“用户端”负责扰动并发送数据样本，“服务器端”负责对数据进行聚合和估计。

请注意，这里的“用户”概念指连接到英特网中的智能设备，如手机、智能手表、汽车和可穿戴设备等，并非持有设备的个人。因此，“用户端”也称为“设备端”。

4.1 单属性态下的 RCP 算法协议

4.1.1 基线算法协议 RCP_Single

首先，在尺度区间 $\left[\frac{1}{e^{\varepsilon}+1},\frac{e^{\varepsilon}}{e^{\varepsilon}+1}\right]$ 上构造一个基线算法协议 RCP_Single，该算法协议与 Duchi 等提出的极值扰动方法（下面简记为 Duchi-EVP）、Wang 提出的分段机制 PM、Nguyen 等提出的 Harmony 方法原理相同，都是基于对称尺度区间的扰动算法，因此估计误差相近。

RCP_Single 构建方法如下：

给定隐私预算 $\varepsilon>0$。在用户端针对每个用户取值 x_i 构建 Bernoulli 随机化器如下：

$$\mathcal{M}_i\sim B\left(1,\frac{(e^{\varepsilon}-1)x_i+1}{e^{\varepsilon}+1}\right)$$

提取 $\mathcal{M}_i$ 的一个取值作为 x_i 的扰动值，即 $\mathcal{M}_i(x_i)=1$ 或 0，将其发送给服务器端；服务器端用聚合器 $\mathcal{Z}$ 将 $\mathcal{M}_i(x_i)$ 聚合为 Z，并利用式(9)计算 x_i 的无偏估计。

算法 1 给出了 RCP_Single 的上层协议。完整的协议由用户端算法 RCP_Single_Client 与服务器端算法 RCP_Single_Server 构成。

RCP_Single_Client 确保离开用户设备的数据具有 ε-差分隐私性，RCP_Single_Server 用于聚合隐私数据，并对均值进行估计。

用户端使用的随机化器 $\mathcal{M}_i$ 和服务器端使用的聚合器 $\mathcal{Z}$ 都是算法协议的基本构建块，我们可以将输入为用户数据 $x_i\in\mathcal{D}$、输出为 x_i 的均值 $\mu\in\mathbb{R}$ 的算法管线看作 $\mathcal{D}$ 到 $\mathbb{R}$ 的“线性映射” $f:\mathcal{D}\to\mathbb{R}$，构成管线的这一族算法被称为神谕器（Oracle），目前神谕器已被广泛用作复杂查询和应用程序的构建块。

算法 1 RCP_Single

输入：用户数据 $x_1,\cdots,x_n$；隐私预算 ε

输出：均值估计值 $\hat{\mu}$

1. for i∈[n] do
2. $\quad\omega_i\leftarrow$RCP_Single_Client(x_i,ε)
3. $\hat{\mu}\leftarrow$RCP_Single_Server(($\omega_1,\cdots,\omega_n$),ε)
4. return $\hat{\mu}$

其中，用户端算法 RCP_Single_Client 的实现如算法 2 所示。

算法 2 RCP_Single_Client

输入：用户 u_i 的数据 x_i；隐私预算 ε

输出：扰动值 ω_i

1. 通过下面的随机化器 $\mathcal{M}_i$ 计算扰动值 ω_i：

$$\omega_i=\begin{cases}1, & \text{w. p. } \dfrac{(e^{\varepsilon}-1)\cdot x_i+1}{e^{\varepsilon}+1}\\ 0, & \text{w. p. } \dfrac{e^{\varepsilon}-(e^{\varepsilon}-1)\cdot x_i}{e^{\varepsilon}+1}\end{cases}$$

2. return ω_i

将服务器端的聚合器定义为一个计数向量 $\boldsymbol{Z}$，初始值为 0，其维数即为属性的维数 m，并将每份客户端报告的 ω_i 累加到 $\boldsymbol{Z}$ 上。这里的 $\boldsymbol{Z}$ 即是式(4)中的 $\frac{1}{n}\sum_{i=1}^{n}1_{x_i\geqslant y_i}$。服务器端算法 RCP_Single_Server 的定义如算法 3 所示。

算法 3 RCP_Single_Server

输入：扰动值序列($\omega_1,\cdots,\omega_n$)；隐私预算 ε

输出：均值估计值 $\hat{\mu}$

1. 初始化 $\mathbf{Z}=0$
2. for i∈[n] do
3. $\quad\mathbf{Z}\leftarrow\mathbf{Z}+\omega_i$
4. $\quad\mathbf{Z}\leftarrow\mathbf{Z}/n$
5. 均值神谕的无偏估计 $\hat{f}(x_i)$ 如下：

$$\hat{f}(x_i)=\frac{(e^{\varepsilon}+1)\cdot\mathbf{Z}-1}{e^{\varepsilon}-1}$$

6. $\quad\hat{\mu}\leftarrow\hat{f}(x_i)$
7. return $\hat{\mu}$

均值神谕 $f:\mathcal{D}\to\mathbb{R}$ 给出了属性 A 的真实均值 μ，即 $f(x_i)=\mu$，但此神谕器（Oracle）是理论上的，是未知的。因此，我们在算法 3 中构建了一个近似的神谕 $\hat{f}$，其输出 $\hat{f}(x_i)$ 是 $f(x_i)$ 的无偏估计，我们将其记为 $\hat{\mu}$ 并输出。

4.1.2 RCP 算法协议的特性

由第 3 节关于随机截尾机制的描述可知，RCP 机制构建的估计算法一定满足 ε-差分隐私性和无偏性。

定理 1（RCP_Single 的效用保证） 设 $\hat{\mu}$ 是 RCP_Single 中用户数据均值的无偏估计，$\hat{\mu}$ 的效用保证以 $\hat{\mu}$ 的方差 $Var(\hat{\mu})$ 来描述，则：

$$Var(\hat{\mu})\leqslant\mathcal{G}_1=\frac{\hat{\mu}-\hat{\mu}^2}{n}+G_1(\varepsilon)\tag{12}$$

其中，$G_1(\varepsilon)=\frac{e^{\varepsilon}}{n(e^{\varepsilon}-1)^2}$ 是误差项，表示差分隐私噪声带来的误差。

证明：根据式(10)：

$$\hat{Var}(\hat{\mu})\leqslant\frac{(d\hat{\mu}+c)-(d\hat{\mu}+c)^2}{d^2n}=\frac{1-2c}{nd}\hat{\mu}-\frac{\hat{\mu}^2}{n}+\frac{c}{nd^2}-\frac{c^2}{nd^2}\tag{13}$$

将 $c=\frac{1}{e^{\varepsilon}+1}$，$d=\frac{e^{\varepsilon}-1}{e^{\varepsilon}+1}$ 代入式(13)即可。

由 $\left[\frac{1}{e^{\varepsilon}+1},\frac{e^{\varepsilon}}{e^{\varepsilon}+1}\right]$ 可以看出，当两端点以不同速率向前运动时，其比值可能保持不变。基于这个直觉，我们在两边同时添加不相等的增量，将其重构为 $\left[\frac{1+\delta_2}{e^{\varepsilon}+1},\frac{e^{\varepsilon}+\delta_1}{e^{\varepsilon}+1}\right]$。这样就有可能达到在不改变隐私保证的基础上优化方差界限 $Var(\hat{\mu})$ 的目的。

定理 2（效用优化定理） 给定一个用于均值估计的 RCP

算法协议和取值范围为[0,1]的输入数据集，设 $\varepsilon>0$ 为系统中给定的隐私预算，对于(0,1]上的任意一点 δ_1，都可以在隐私保证 ε 不改变的条件下，利用 $\delta_2\in[\delta_1/e^{\varepsilon},\delta_1]$ 将基线协议的尺度区间 $\left[\frac{1}{e^{\varepsilon}+1},\frac{e^{\varepsilon}}{e^{\varepsilon}+1}\right]$ 重构为 $\left[\frac{1+\delta_2}{e^{\varepsilon}+1},\frac{e^{\varepsilon}+\delta_1}{e^{\varepsilon}+1}\right]$，使得新的算法协议具有更优的均值估计方差界限。且：

(1)δ_1 越大，构造出的算法效用越大；

(2)当固定 δ_1 时，δ_2 在$[\delta_1/e^{\varepsilon},\delta_1]$区间内取值。算法的效用随 δ_2 的减小而增大，当 $\delta_2=\delta_1/e^{\varepsilon}$ 时，效用取得极大值。

(3)差分隐私状态越高(ε 越小)，调节 δ_1 的取值所取得的效果就越显著。

证明：1)任取 $\delta_1,\delta_2,\delta_3\in(0,1]$ 且 $\frac{\delta_1}{e^{\varepsilon}}\leqslant\delta_2\leqslant\delta_1$，$\delta_3=\delta_1-\delta_2$，令 $c=\frac{1+\delta_2}{e^{\varepsilon}+1}$，$c+d=\frac{e^{\varepsilon}+\delta_1}{e^{\varepsilon}+1}$，则由式(11)可得：

$$\frac{c+d}{c}=\frac{e^{\varepsilon}+\delta_1}{1+\delta_2}\leqslant e^{\varepsilon}$$

因此，尺度区间为 $\left[\frac{1+\delta_2}{e^{\varepsilon}+1},\frac{e^{\varepsilon}+\delta_1}{e^{\varepsilon}+1}\right]$ 的算法协议仍满足 ε-差分隐私。将 c 与 d 代入式(13)：

$$\hat{Var}(\hat{\mu})\leqslant\frac{1}{n}\left(\frac{e^{\varepsilon}-1-2\delta_2}{e^{\varepsilon}-1+\delta_3}\hat{\mu}-\hat{\mu}^2\right)+\frac{1}{n}\frac{(e^{\varepsilon}+1)(1+\delta_2)}{(e^{\varepsilon}-1+\delta_3)^2}-\frac{1}{n}\left(\frac{1+\delta_2}{e^{\varepsilon}-1+\delta_3}\right)^2$$

该条件下的方差界限 $\mathcal{G}_2$ 如下：

$$\mathcal{G}_2(\delta_1,\delta_2)=\frac{1}{n}\frac{e^{\varepsilon}-1-2\delta_2}{e^{\varepsilon}-1+\delta_1-\delta_2}\hat{\mu}-\frac{\hat{\mu}^2}{n}+\frac{1}{n}\frac{(e^{\varepsilon}+1)(1+\delta_2)}{(e^{\varepsilon}-1+\delta_1-\delta_2)^2}-\frac{1}{n}\frac{(1+\delta_2)^2}{(e^{\varepsilon}-1+\delta_1-\delta_2)^2}$$

为简单起见，将其记为：

$$\mathcal{G}_2(\delta_1,\delta_2)=\frac{1}{n}\cdot g_{21}(\delta_1,\delta_2)\cdot\hat{\mu}-\frac{\hat{\mu}^2}{n}+\frac{1}{n}\cdot g_{22}(\delta_1,\delta_2)-\frac{1}{n}\cdot g_{23}(\delta_1,\delta_2)\tag{14}$$

如果假设：

$$G_2(\varepsilon,\delta_2,\delta_3)=\frac{(e^{\varepsilon}+1)(1+\delta_2)}{n(e^{\varepsilon}-1+\delta_3)^2}-\frac{(1+\delta_2)^2}{n(e^{\varepsilon}-1+\delta_3)^2}$$

则式(14)也可记为：

$$\mathcal{G}_2=\frac{1}{n}\frac{e^{\varepsilon}-1-2\delta_2}{e^{\varepsilon}-1+\delta_3}\hat{\mu}-\frac{\hat{\mu}^2}{n}+G_2(\varepsilon,\delta_2,\delta_3)\tag{15}$$

其中，$G_2(\varepsilon,\delta_2,\delta_3)$是误差项。

2)下面证明对于 $\forall\delta_1\in(0,1]$，一定存在 $\delta_2\in\left[\frac{\delta_1}{e^{\varepsilon}},\delta_1\right]$，使得 $\mathcal{G}_2<\mathcal{G}_1$。

固定式(14)中的 δ_1，对 δ_2 求导，由于式中 $0\leqslant\hat{\mu}\leqslant1$，因此：

$$\frac{\partial\mathcal{G}_2}{\partial\delta_2}=\frac{2(1-\delta_1)(e^{\varepsilon}+\delta_1)}{n(e^{\varepsilon}-1+\delta_1-\delta_2)^3}>0$$

因此，$\mathcal{G}_2$ 关于 δ_2 单增，$\frac{\delta_1}{e^{\varepsilon}}$ 为 $\mathcal{G}_2$ 的极小值点，将 $\delta_2=\frac{\delta_1}{e^{\varepsilon}}$，$\delta_3=\frac{e^{\varepsilon}-1}{e^{\varepsilon}}\delta_1$ 代入 $\mathcal{G}_2$ 得：

$$\mathcal{G}_2<\frac{1}{n}\hat{\mu}-\frac{\hat{\mu}^2}{n}+\frac{(e^{\varepsilon}+1)}{n(e^{\varepsilon}-1)^2}\cdot\frac{e^{\varepsilon}}{e^{\varepsilon}+\delta_1}-\frac{1}{n(e^{\varepsilon}-1)^2}<\frac{1}{n}\hat{\mu}-\frac{\hat{\mu}^2}{n}+\frac{e^{\varepsilon}}{n(e^{\varepsilon}-1)^2}=\mathcal{G}_1$$

由 δ_1 的任意性可知，无论 δ_1 在(0,1]如何取值，使得 $\mathcal{G}_2<\mathcal{G}_1$ 的 δ_2 一定存在。因此，尺度区间为 $\left[\frac{1+\delta_2}{e^{\varepsilon}+1},\frac{e^{\varepsilon}+\delta_1}{e^{\varepsilon}+1}\right]$ 的算法协议比基线协议拥有更优的均值估计方差界限。

3)考察定理附加的推论(1)－(3)：(2)是显然的；关于(3)，我们会在推论1的证明中详细论述。下面我们证明(1)：

固定式(14)中的 δ_2，对 δ_1 求导：

$$\frac{\partial\mathcal{G}_2}{\partial\delta_1}=\frac{1}{n}\cdot\frac{\partial g_{21}}{\partial\delta_1}\cdot\hat{\mu}+\frac{1}{n}\cdot\frac{\partial g_{22}}{\partial\delta_1}-\frac{1}{n}\cdot\frac{\partial g_{23}}{\partial\delta_1}$$

由于 $\frac{\partial\mathcal{G}_2}{\partial\delta_1}$ 是关于 $\hat{\mu}$ 的线性函数，因此我们分别将 $\hat{\mu}=0$ 和 $\hat{\mu}=1$ 代入上式得：

$$\frac{\partial\mathcal{G}_2}{\partial\delta_1}\Big|_{\hat{\mu}=0}=-\frac{2(1+\delta_2)(e^{\varepsilon}-\delta_2)}{n(e^{\varepsilon}-1+\delta_1-\delta_2)}<0$$

$$\frac{\partial\mathcal{G}_2}{\partial\delta_1}\Big|_{\hat{\mu}=1}=-\frac{(e^{\varepsilon}+1)(e^{\varepsilon}-1+\delta_1-\delta_2)}{n(e^{\varepsilon}-1+\delta_1-\delta_2)}-\frac{2(1+\delta_2)(1-\delta_1)}{n(e^{\varepsilon}-1+\delta_1-\delta_2)}<0$$

因此，$\mathcal{G}_2$ 关于 δ_1 单调递减，即如(1)中的描述：随着 δ_1 的增大，$\hat{\mu}$ 的方差界限减小，算法效用提高。

由定理1很容易得到推论1。

推论1 RCP_Single 中的 $\hat{\mu}$ 估计方差的最优界限 $\mathcal{G}_{\text{opt}}$ 在 $\delta_1=1,\delta_2=1/e^{\varepsilon}$ 处取得：

$$\mathcal{G}_{\text{opt}}=\frac{1}{n}\left(\frac{e^{\varepsilon}-2}{e^{\varepsilon}-1}\hat{\mu}-\hat{\mu}^2\right)+G_{\text{opt}}(\varepsilon)\tag{16}$$

其中，$G_{\text{opt}}(\varepsilon)=\frac{1}{n(e^{\varepsilon}-1)}$，此时尺度区间参数为：$c=\frac{1}{e^{\varepsilon}}$，$d=\frac{e^{\varepsilon}-1}{e^{\varepsilon}}$。

证明：根据定理2，显然有 $\mathcal{G}_2$ 随 δ_1 增大而减小，随 δ_2 增大而增大，因此在 $\delta_1=1,\delta_2=\frac{1}{e^{\varepsilon}}$ 时取得极小值，将 $\delta_1=1,\delta_2=\frac{1}{e^{\varepsilon}}$ 代入式(14)即得式(16)。

推论1给出了算法协议 RCP_Single 的效用优化公式。

将式(16)与式(12)进行比较并计算：

$$\mathcal{G}_1-\mathcal{G}_{\text{opt}}=\frac{\hat{\mu}}{n}\frac{1}{e^{\varepsilon}-1}+\frac{1}{n(e^{\varepsilon}-1)^2}$$

不妨令 $\hat{\mu}=1$，考察曲线 $\frac{e^{\varepsilon}}{n(e^{\varepsilon}-1)^2}$。根据 $\frac{e^{\varepsilon}}{(e^{\varepsilon}-1)^2}$ 下的凸曲线特征：当 $\varepsilon\leqslant1$，即高隐私态时，随 ε 的减小，$\mathcal{G}_1-\mathcal{G}_{\text{opt}}$ 增长显著，即算法在高隐私态优势明显，这与定理2的附加推论(3)的描述相符；另一方面，由于 $\frac{e^{\varepsilon}}{(e^{\varepsilon}-1)^2}$ 下凸曲线拐点在1～4之间，因此当 $\varepsilon>1$，即低隐私态以后，$\mathcal{G}_1$ 与 $\mathcal{G}_{\text{opt}}$ 的差距迅速趋近于0并形成拖尾。因此，在低隐私态下，RCP 算法与同类算法相比并不具备明显优势。从后面的实验中可以看到，与对低隐私态友好的 PM 算法相比，RCP 在 $\varepsilon>1$ 时性能并不会特别占优。

我们知道，当 $c=0,d=1$ 时，RCP_Single 将退化为图2中的经典随机截尾模型，因此失去 ε-差分隐私性，我们将这个退化的 RCP 算法协议称为“纯”RCP 算法协议。

很容易看出,“纯”RCP 算法协议的精度高于其他任何 RCP 算法协议。因此,可以引出下面均值估计 $\hat{\mu}$ 的方差下界定理。

定理 3(方差下界定理) RCP_Single 对应的“纯”RCP 算法协议的 $\hat{\mu}$ 估计方差界限为:

$$Var(\hat{\mu}) \geqslant \mathcal{G}_{\text{pure}} = \frac{\mu - \hat{\mu}^2}{n} \tag{17}$$

且 $\mathcal{G}_{\text{pure}}$ 是所有基于 RCP 机制的单属性差分隐私算法系统中均值 $\hat{\mu}$ 估计方差的下界。

4.2 多属性态下的 RCP 算法协议

在多属性态下我们会采用类似于计数均值草图(Count Mean Sketch)算法的公币模型(Public Coin Model)技术[21],使得用户只需要向服务器端发送 1 位,在保证相同的精度级别的基础上大大降低了通信成本。

4.2.1 基线算法协议 RCP_Multiple

我们同样在对称尺度区间 $\left[\frac{1}{e^{\varepsilon}+1}, \frac{e^{\varepsilon}}{e^{\varepsilon}+1}\right]$ 上构造了一个基线算法协议 RCP_Multiple。设用户 u_i 的数据样本为 $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{im})$,均值神谕(Oracle)为 $f: \mathcal{D}^m \to \mathbb{R}^m$,且输出 $f(x_{\cdot j})(j=1, \cdots, m)$ 等于 $x_{\cdot j}$ 的真实均值 μ_j,$\hat{f}(x_{\cdot j}) = \hat{\mu}_j$ 是 $f(x_{\cdot j})$ 的无偏估计,均值向量估计 $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \cdots, \hat{\mu}_m)$ 是协议的最终输出结果。

从 RCP_Multiple 的上层协议(见算法 4)来看,完整协议仍由用户端与服务器端两部分组成,不同的是用户端引入一个均匀随机化器 $\mathcal{R}_0$,即公币随机性。用于随机均匀地获取 $[m] = \{1, \cdots, m\}$ 中的一个索引值,利用该值确保在合理精度下的数据通信量由 m 下降到 1。

算法 4 RCP_Multiple

输入:用户数据 $x_1, \cdots, x_n$;隐私预算 ε

输出:均值向量估计值 $\hat{\boldsymbol{\mu}}$

1. for i∈[n] do
2. $\langle \omega_i, j^{(i)} \rangle \leftarrow$RCP_Multiple_Client($x_i, \varepsilon$)
3. Z←RCP_Aggregator($\langle \omega_1, j^{(1)} \rangle, \cdots, \langle \omega_n, j^{(n)} \rangle$)
4. for j∈[m] do
5. $\hat{\mu}_j \leftarrow$RCP_Multiple_Server(Z, j, ε)
6. $\hat{\boldsymbol{\mu}} \leftarrow (\hat{\mu}_1, \cdots, \hat{\mu}_m)$
7. return $\hat{\boldsymbol{\mu}}$

算法 5 RCP_Multiple_Client

输入:用户 u_i 的数据 x_i;隐私预算 ε

输出:扰动值对 $\langle \omega_i, j^{(i)} \rangle$

1. 利用 $\mathcal{R}_0$ 随机均匀抽取样本 $j^{(i)} \leftarrow [m]$
2. 利用随机化器 $\mathcal{M}_i$ 计算扰动值 ω_i:

$$\omega_i = \begin{cases} 1, & \text{w. p. } \dfrac{(e^{\varepsilon}-1) \cdot x_{i,j^{(i)}} + 1}{e^{\varepsilon}+1} \\ 0, & \text{w. p } \dfrac{e^{\varepsilon} - (e^{\varepsilon}-1) \cdot x_{i,j^{(i)}}}{e^{\varepsilon}+1} \end{cases}$$

3. return $\omega_i, j^{(i)}$

我们为服务器端 RCP_Multiple_Server 单独封装聚合器,如算法 6 所示。

算法 6 RCP_Aggregator

输入:n 个扰动值对 $\langle \omega_1, j^{(1)} \rangle, \cdots, \langle \omega_n, j^{(n)} \rangle$

输出:计数向量 $\mathbf{Z}$

1. $\mathbf{Z} \leftarrow \{0\}^m$
2. for i∈[n] do $\mathbf{Z}[j^{(i)}] \leftarrow \mathbf{Z}[j^{(i)}] + m \cdot \omega_i$
3. $\mathbf{Z} \leftarrow \mathbf{Z}/n$
4. return $\mathbf{Z}$

服务器端算法 RCP_Multiple_Server 的处理过程如算法 7 所示。

算法 7 RCP_Multiple_Server

输入:计数向量 $\mathbf{Z}$;$\mathbf{Z}$ 的索引 j;隐私预算 ε

输出:均值估计值 $\hat{\mu}_j$

1. 均值神谕的无偏估计 $\hat{f}(x_{\cdot j})$ 如下:

$$\hat{f}(x_{\cdot j}) = \frac{(e^{\varepsilon}+1) \cdot \mathbf{Z}[j] - 1}{e^{\varepsilon}-1}$$

2. $\hat{\mu}_j \leftarrow \hat{f}(x_{\cdot j})$
3. return $\hat{\mu}_j$

可以看出,RCP_Multiple 仍然满足 ε-差分隐私且估计值无偏,$\hat{\mu}_j$ 估计方差的界限如下:

$$\hat{Var}(\hat{\mu}_j) \leqslant \mathcal{G}_3 = \frac{1}{n}\left(\frac{m(e^{\varepsilon}+1)-2}{(e^{\varepsilon}-1)}\hat{\mu}_j - \hat{\mu}_j^2\right) + G_3(\varepsilon)$$

其中,$G_3(\varepsilon) = \frac{m}{n} \cdot \frac{e^{\varepsilon}+1}{(e^{\varepsilon}-1)^2} - \frac{1}{n(e^{\varepsilon}-1)^2}$,表示隐私预算 ε 引起的误差。

显然,$G_3(\varepsilon) \geqslant G_1(\varepsilon)$,当 $m=1$ 时二者相等。这是由于引入了 $\mathcal{R}_0$ 的公币随机性,使得多属性协议接收了更多的噪声,因此误差比单属性协议大,但通信效率会比不施加 $\mathcal{R}_0$ 更高。

从公式上看:$\mathcal{G}_3 \geqslant \mathcal{G}1 \geqslant \mathcal{G}_{\text{opt}}$。由定理 2 可知,随着 ε 逐渐增大,$\mathcal{G}_3$ 和 $\mathcal{G}_1$ 会收敛于 $\mathcal{G}_{\text{opt}}$。

4.2.2 多属性态下的优化算法协议

显然,效用优化定理在多属性态下同样成立,我们利用定理的结论可构造出比 RCP_Multiple 估计精度更高的优化算法协议 RCP_Multiple_Opt。

算法 8 RCP_Multiple_Opt

输入:用户数据 $x_1, \cdots, x_n$;隐私预算 ε;优化参数 δ_1, δ_2

输出:均值向量估计值 $\hat{\boldsymbol{\mu}}$

1. for i∈[n] do
2. $\langle \omega_i, j^{(i)} \rangle \leftarrow$RCP_Multiple_Client_Opt($x_i, \varepsilon, \delta_1, \delta_2$)
3. $\mathbf{Z} \leftarrow$RCP_Aggregator($\langle \omega_1, j^{(1)} \rangle, \cdots, \langle \omega_n, j^{(n)} \rangle$)
4. for j∈[m] do
5. $\hat{\mu}_j \leftarrow$RCP_Multiple_Server_Opt($\mathbf{Z}, j, \varepsilon, \delta_1, \delta_2$)
6. $\hat{\boldsymbol{\mu}} \leftarrow (\hat{\mu}_1, \cdots, \hat{\mu}_m)$
7. return $\hat{\boldsymbol{\mu}}$

协议中的上层协议 RCP_Multiple_Opt 的执行步骤同 RCP_Multiple 是一致的。但按照效用优化定理,将用户端算法与服务器端算法进行了修改。其中,用户端算法 RCP_Multiple_Client_Opt 与 RCP_Multiple_Client 相比,将随机化器 M_i 按照尺度区间 $\left[\frac{1+\delta_2}{e^{\varepsilon}+1}, \frac{e^{\varepsilon}+\delta_1}{e^{\varepsilon}+1}\right]$ 更新为:

$$\omega_i = \begin{cases} 1, & \text{w. p. } \dfrac{(e^{\varepsilon}-1+\delta_1-\delta_2) \cdot x_{i,j^{(i)}} + 1 + \delta_2}{e^{\varepsilon}+1} \\ 0, & \text{w. p. } \dfrac{e^{\varepsilon}-\delta_2-(e^{\varepsilon}-1+\delta_1-\delta_2) \cdot x_{i,j^{(i)}}}{e^{\varepsilon}+1} \end{cases}$$

服务器端 RCP_Multiple_Server_Opt 对应地将聚合器 $\mathcal{X}$ 更新为:

$$\hat{f}(x_{\cdot j})=\frac{(e^{\varepsilon}+1)\cdot \boldsymbol{Z}[j]-1-\delta_2}{e^{\varepsilon}-1+\delta_1-\delta_2}$$

在多属性态下，$\hat{\mu}_j$ 的方差最优界限仍然在 $\delta_1=1,\delta_2=1/e^{\varepsilon}$ 处取得：

推论 2 RCP_Multiple 中的 $\hat{\mu}_j$ 估计方差的最优界限 $\mathcal{G}_{\text{opt}}^{(j)}$ 在 $\delta_1=1,\delta_2=1/e^{\varepsilon}$ 处取得：

$$\mathcal{G}_{\text{opt}}^{(j)}=\frac{1}{n}\left(\frac{me^{\varepsilon}-2}{e^{\varepsilon}-1}\hat{\mu}_j-\hat{\mu}_j^2\right)+G_{\text{opt}}^{(j)}(\varepsilon) \tag{18}$$

其中，$G_{\text{opt}}^{(j)}(\varepsilon)=\frac{me^{\varepsilon}-1}{n(e^{\varepsilon}-1)^2}$，此时尺度区间参数为：$c=\frac{1}{e^{\varepsilon}}$，$d=\frac{e^{\varepsilon}-1}{e^{\varepsilon}}$。

我们可以将向量 $\hat{\boldsymbol{\mu}}$ 的方差最优界限记为：

$$\mathcal{G}_{\text{opt}}=(\mathcal{G}_{\text{opt}}^{(1)},\cdots,\mathcal{G}_{\text{opt}}^{(m)}) \tag{19}$$

最后，我们给出多属性态下 RCP 算法协议的误差下界，该界限与单属性态相比是一致的。

定理 4（方差下界定理） RCP_Multiple 对应的“纯”RCP 算法协议的 $\hat{\mu}_j$ 估计方差为：

$$Var(\hat{\mu}_j)\geqslant\mathcal{G}_{\text{pure}}^{(j)}=\frac{\hat{m}\mu_j-\hat{\mu}_j^2}{n} \tag{20}$$

且 $\mathcal{G}_{\text{pure}}^{(j)}$ 是所有基于 RCP 机制的多属性差分隐私算法系统中均值 $\hat{\mu}_j$ 估计方差的下界。

5 实验评估

5.1 实验环境

本次实验通过智能设备采集数据的大规模真实数据集来验证本文算法的有效性和准确性，运行环境为：Windows 10 64 位操作系统，Inter(R) Core(TM) i7-6600U CPU @2.60 GHz 2.81 GHz，24.0 GB 内存，Python 3.9.7，Numpy 1.20.3，Pandas 1.3.4，Scipy 1.7.1，R 4.1.2。

5.2 数据集

我们采用加州大学欧文分校的 UCI 机器学习库中的数据集进行实验，选取了来自公共存储库的智能手机和智能手表的人类活动识别异构数据集（Heterogeneity Dataset for Human Activity Recognition，HHAR）。该数据集可在加州大学欧文分校提供的站点下载[1)]。

该数据集用于测试包含异质性传感器的人类活动识别算法。其中，包含两个设备（智能手机和智能手表）各自的两个运动传感器（加速计和陀螺仪）的读数记录。

用于实验的数据集从智能手机的传感器数据中提取，即以下两个数据文件：1) Phones_accelerometer.csv，13 062 475 条记录；2) Phones_gyroscope.csv，13 932 632 条记录。

文件中数据记录的属性相同，共 10 个属性，如表 1 所列。表 1 中 Gt 共有 6 个取值，即 bike，sit，stand，walk，stairsup，stairsdown，分别代表传感器记录的 6 类活动，每个文件的每类活动各有 200 万条左右的记录。x,y,z 为数值型数据，表示传感器采集的空间坐标。我们将分别提取 accelerometer 和 gyroscope 文件中的 6 类活动的 x,y,z 作为实验数据，通过对 x,y,z 进行均值估计来验证所提出算法的精度。

表 1 HHAR 数据集的属性列表

Table 1 Attributes of HHAR dataset

属性名称	说明
Index	数据记录的行数
Arrival_Time	测量值到达传感器的时间
Creation_Time	附加到样本的时间戳
x	传感器为 *x* 轴提供的值
y	传感器为 *y* 轴提供的值
z	传感器为 z 轴提供的值
User	采样用户（命名为 a 到 i）
Model	样本源自的手机/手表型号
Device	样本来自的特定设备
Gt	用户正在执行的活动

5.3 数据预处理

两个数据文件的列名相同，我们仅以“手机-加速计”数据文件 Phones_accelerometer.csv 为例展示数据内容。

表 2 Phones_accelerometer.csv 的数据

Table 2 Data of Phones_accelerator.csv

Index	Arrival_Time	Creation_Time
0	1424696633908	1424696631913248572
1	1424696633909	1424696631918283972
2	1424696633918	1424696631923288855
⋮	⋮	⋮
13062475	1424778553395	92263861839000

x	*y*	*z*	User
−5.958191	0.688065	8.135345	a
−5.952240	0.670212	8.136536	a
−5.995087	0.653549	8.204376	a
⋮	⋮		
1.379043	0.0	9.959755	i

Model	Device	Gt
nexus4	nexus4_1	stand
nexus4	nexus4_1	stand
nexus4	nexus4_1	stand
⋮	⋮	⋮
samsungold	samsungold_2	bike

我们分别提取 accelerometer 和 gyroscope 文件的 6 类活动 bike，sit，stand，walk，stairsup，stairsdown 的 x,y,z 作为实验数据，因此实验数据集的维数 $n=2\times6\times3=36$，为避免重名，在生成实验数据集时将 x,y,z 的列名按传感器和 Gt 类型重命名，如表 3 所列。

表 3 实验数据集的属性列表

Table 3 Attributes of experimental dataset

accelerometer			
bike	sit	stand	walk
p_a_bike_x	p_a_sit_x	p_a_std_x	p_a_walk_x
p_a_bike_y	p_a_sit_y	p_a_std_y	p_a_walk_y
p_a_bike_z	p_a_sit_z	p_a_std_z	p_a_walk_z

accelerometer		gyroscope	
stairsup	stairsdown	bike	sit
p_a_up_x	p_a_down_x	p_g_bike_x	p_g_sit_x
p_a_up_y	p_a_down_y	p_g_bike_y	p_g_sit_y
p_a_up_z	p_a_down_z	p_g_bike_z	p_g_sit_z

gyroscope			
stand	walk	stairsup	stairsdown
p_g_std_x	p_g_walk_x	p_g_up_x	p_g_down_x
p_g_std_y	p_g_walk_y	p_g_up_y	p_g_down_y
p_g_std_z	p_g_walk_z	p_g_up_z	p_g_down_z

1) http://archive.ics.uci.edu/ml/datasets/Heterogeneity+Activity+Recognition

如表 3 所列，实验数据集的数据记录共有 36 个属性，每一列属性都代表某个设备的特定传感器的特定活动记录，例如 p_g_walk_x 表示手机(Phone)的陀螺仪(Gyroscope)所记录 walk 活动的 x 轴坐标数值。

具体抽样过程如下：

将样本量 n 设为 800 000，分别在“手机-加速计”文件 Phones_accelerometer.csv 和“手机-陀螺仪”文件 Phones_gyroscope.csv 中按 6 个 Gt 属性随机均匀抽取 80 万行 x,y,z 列的数据子表，再将这 12 个子表按上述要求修改列名，按列合并，然后对数据进行归一化，最终得到一个待测试的实验数据集。

按上述步骤重复进行抽样，最后生成 30 个实验数据集作为一组测试用例，并且编号为：1，…，30。

实验中 RCP 算法将与 Duchi 等的算法进行计较。由于 Duchi 等的算法的输入域为[-1,1]，为了方便比较，我们将每个实验数据集分别归一化到[0,1]和[-1,1]。我们在[0,1]数据集上运行 RCP 算法；在[-1,1]数据集上运行 Duchi 等其他算法，并将得到的[-1,1]区间上的均值向量估计值归一化到[0,1]区间。

5.4 实验设计

本实验通过在多属性状态下，RCP 机制中的效用优化算法 RCP_Multiple_Opt 与基线算法 RCP_Multiple、Duchi 的极值扰动(Duchi-EVP)算法、Wang 的 PM 算法、Nguyen 的 Harmony 算法的比较，展现在相同的软硬件环境、数据样本规模和隐私预算条件下 RCP 机制的均值估计能力[22-27]。

这里以估计精度(即效用)作为算法性能的度量。实验中选取均方根误差(Root Mean Squared Error，RMSE)作为评价指标。

假设实验数据集的数据有 m 个属性，真实均值向量为 $\boldsymbol{\mu}=(\mu_1,\cdots,\mu_m)$，均值向量估计为 $\hat{\boldsymbol{\mu}}=(\hat{\mu}_1,\cdots,\hat{\mu}_m)$，则算法的均方根误差为：

$$RMSE=\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\hat{\mu}_j-\mu_j)^2} \tag{21}$$

在实验开始前计算每个实验数据集的每一列的均值得到 $\boldsymbol{\mu}$。

对于每种算法，我们在客户端采用相同的公币随机化器 $\mathcal{R}_0$。即对于每条数据，仅随机均匀选取一个属性项扰动并发送到服务器端。

我们将会测试不同级别的隐私预算条件下算法的执行效果：

$\varepsilon=0.1,0.2,0.5,0.8,1,2,5,8,10$

给定 ε，依次在 1，…，30 号实验数据集上分别运行待检验的 5 种算法。

我们在不同 ε 条件下对测量结果分别运用 Friedman 秩相关分析法，通过秩(Rank)来检验算法的性能是否存在显著性差异。

Friedman 的假设检验问题如下。H_0 表示样本间无显著差异；H_1 表示样本间有显著差异。假设有 k 个待检验的算法(即样本)在 l 个数据集上进行测试，则将 k 种算法在每个数据集上按 RMSE 从小到大排序，并在每种算法上标注排序编号 I_{ij}，即对指标进行评秩，其中 $1\leqslant j\leqslant k$，$1\leqslant i\leqslant l$，$1\leqslant I_{ij}\leqslant k$。编号 1 的算法在该数据集上效用最优，反之编号 k 的效用最差。算法的平均排序值(均秩)和检验统计量如下：

$$\bar{I}_j=\frac{1}{l}\sum_{i=1}^{l}I_{ij} \tag{22}$$

$$\chi_F^2=\frac{12l}{k(k+1)}\left(\sum_{j=1}^{k}\bar{I}_j^2-\frac{k(k+1)^2}{4}\right) \tag{23}$$

当 k 固定，l 足够大时，H_0 下 χ_F^2 近似服从自由度为 $k-1$ 的 χ^2 分布。

5.5 实验结果与分析

5.5.1 $\varepsilon=0.1$ 时的 Friedman 检验举例

为了方便排名，我们将算法作为列向量，将实验数据集作为行向量，按性能指标分别列出每种算法在不同数据集上的 RMSE 以及 RMSE 的平均值。

当 $\varepsilon=0.1$ 时，分别在 1-30 号实验数据集上执行 RCP_Multiple_Opt 等 5 种算法，得到其均值估计量，然后利用式(21)计算出各协议在各数据集上的 RMSE 指标，如表 4 所列。

表 4 $\varepsilon=0.1$ 时 RCP_Multiple_Opt 等 5 种算法在 1-30 号实验数据集上均值估计的 RMSE 指标对比

Table 4 Comparison of RMSE indicators of the mean estimated value of 5 algorithms such as RCP_Multiple_Opt on test datasets 1-30 when $\varepsilon=0.1$

Dataset Index	RCP_Multiple_Opt	RCP_Multiple	Duchi-EVP	PM	Harmony
1	0.059470	0.088126	0.094008	0.065735	0.088181
2	0.060536	0.093125	0.102180	0.072534	0.075326
3	0.058431	0.087182	0.091314	0.057236	0.096142
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
28	0.059143	0.061416	0.089816	0.072175	0.091512
29	0.063140	0.103135	0.082037	0.069048	0.112139
30	0.065233	0.098432	0.068138	0.072126	0.058325
RMSE mean	0.060946	0.087023	0.084699	0.071142	0.081232

从表 4 可以看出，RCP_Multiple_Opt 在绝大部分数据集上表现最好，优势明显，其次为 PM 算法；而 RCP_Multiple，Duchi-EVP 和 Harmony 算法的表现相对较差。

下文利用 Friedman 检验判断上述算法的性能在 $\varepsilon=0.1$ 时是否具有显著差异，这里我们将显著性水平 α 设为 0.01。

将表 4 所列的测量数据转化为定序尺度。按 RMSE 性能指标，将 5 种算法在所有实验数据集上分别从小到大进行排名，每种算法上标注排名编号 I_{ij}(秩)。每种算法在 $\varepsilon=0.1$ 时在各数据集上的排名和平均排名(均秩)如表 5 所列。

表 5 $\varepsilon=0.1$ 时 RCP_Multiple_Opt 等 5 种算法在 1－30 号实验数据集上 RMSE 指标的排名

Table 5 Ranking of RMSE indicators of 5 algorithms such as RCP_Multiple_Opt on test data sets 1－30 when $\varepsilon=0.1$

Dataset Index	RCP_Multiple_Opt	RCP_Multiple	Duchi-EVP	PM	Harmony
1	1	3	5	2	4
2	1	4	5	2	3
3	2	3	4	1	5
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮
28	1	2	4	3	5
29	1	4	3	2	5
30	2	5	3	4	1
mean rank	1.4	3.833333	3.7	2.4	3.666667

将 $k=5$，$l=30$ 和均秩 $\bar{I}_j$ 代入式(23)，得到 Friedman 统计量为 54.587，对应 p 值为 3.966×10^{-11}，远小于 α，因此算法之间性能差异明显，且 RCP_Multiple_Opt 性能的平均排名高于其他算法。

5.5.2 不同 ε 条件下基于秩的性能分析

以同样的方式获取各算法在 $\varepsilon=0.2$ 到 $\varepsilon=10.0$ 时的指标排名结果和 Friedman 检验结果，如表 6 所列。

表 6 不同的 ε 下 RCP_Multiple_Opt 等 5 种算法的均秩列表

Table 6 Mean rank of 5 algorithms such as RCP_Multiple_Opt with different ε

ε	RCP_Multiple_Opt	RCP_Multiple	Duchi-EVP	PM	Harmony
0.1	1.40	3.83	3.70	2.40	3.67
0.2	1.37	3.90	3.77	2.67	3.70
0.5	1.44	3.78	3.81	2.37	3.59
0.8	1.47	3.90	3.83	2.47	3.37
1.0	1.60	3.83	3.63	2.50	3.43
2.0	1.80	3.60	3.47	2.86	3.27
5.0	2.13	3.50	3.33	2.80	3.23
8.0	2.67	3.43	3.53	2.57	3.20
10.0	2.23	3.40	3.67	2.73	3.27

表 6 中每一行数据为给定 ε 条件下，5 种算法在所有实验数据集上执行后，通过式(22)得到的性能指标的平均排序值 $\bar{I}_j$。如果算法之间的性能无显著差异，秩 I_{ij} 的分布应该是随机的，即每种算法的均秩 $\bar{I}_j$ 应大致相等；反之，$\bar{I}_j$ 值之间会有明显不同。

从表 6 可以看出，RCP_Multiple_Opt 和 PM 在 $\varepsilon<0.1$（高隐私状态）时均秩 $\bar{I}_j$ 明显小于其他算法，且 RCP_Multiple_Opt 的均秩比 PM 平均小 1 左右。这是由于 RCP_Multiple_Opt 遵循效用优化定理对对称尺度区间 $\left[\frac{1}{e^{\varepsilon}+1},\frac{e^{\varepsilon}}{e^{\varepsilon}+1}\right]$ 进行了优化，估计方差达到最优界限 $\mathscr{G}_{\text{opt}}$ 后会比 RCP_Multiple 小很多，因此排名的表现最好。而 Duchi-EVP 和 Harmony 都采用对称尺度区间，设计思想与 RCP_Multiple 相似，误差也必然相近。

Wang 的 PM 属于分布扰动机制，通过定义[－1，1]更宽的域[－s，s]来输出不同概率的连续值，比 Duchi 等的方差小，因此平均排序值比 Duchi-EVP，Harmony 和 RCP_Multiple 更小。

此外，$\varepsilon<0.1$ 时，RCP_Multiple_Opt 的均秩比其他算法小很多，这与定理 1 中描述的“隐私状态越高，优化效果越显著”的结论相符。

在 $\varepsilon=1$（中隐私状态）以后，5 种算法之间的差距逐渐缩小。但在 $\varepsilon=2$ 和 $\varepsilon=5$ 时，RCP_Multiple_Opt 的均秩仍然保持第一，其取值缓慢上升。Duchi-EVP，Harmony 和 RCP_Multiple 的均秩迅速下降，开始与 Opt 算法接近。PM 则保持相对稳定，这说明 $\varepsilon>1$ 时，PM 更具优势。

当 $\varepsilon=8$ 和 $\varepsilon=10$ 时，RCP_Multiple_Opt 的平均排名与 PM 接近甚至有时被超过，这说明 RCP_Multiple_Opt 在低隐私态（$\varepsilon>1$）下的表现不如高隐私态，这一点从推论 1 和推论 2 即可看出，在 $\varepsilon>1$ 后，RCP_Multiple_Opt 的方差界限会迅速接近最优界限 $\mathscr{G}_{\text{opt}}$，难以与其他算法拉开差距。

此外，Wang 的 PM 采用的输出域[－s，s]比[－1，1]更宽，且以高概率输出[－s，s]中心邻域的值，以低概率输出[－s，s]两端邻域的值。该方案已被证明得到的估计值是均值的无偏估计且在一定程度上弥补了 Duchi 方案的缺陷。因此，当 ε 较大时，PM 与 Duchi-EVP 相比方差会小很多；而此时 PM 与 RCP_Multiple_Opt 相比，尽管经过优化后 RCP_Multiple_Opt 更接近 $\mathscr{G}_{\text{opt}}$，但一方面低隐私态下 RCP_Multiple_Opt 性能变化趋缓，另一方面 PM 针对低隐私态性能有较大提升，因此 PM 在准确性上仍具有一定优势。

RCP_Multiple_Opt 在低隐私态下与 Duchi-EVP，Harmony 和 RCP_Multiple 相比仍具有明显优势。

在低隐私态下，Duchi-EVP，Harmony 和 RCP_Multiple 的均秩取值已经相当接近。而同高隐私态相比，它们与 PM 和 RCP_Multiple_Opt 的差距也缩小了不少。这是由于 ε 越大，式(19)中由 ε 带来的误差就会越小。根据定理 3，所有 RCP 类别的估计方差都会收敛到下界 $\mathscr{G}^{(1)}_{\text{pure}}$。

表 7 中，每一行为给定 ε 条件下，对评秩后的定序尺度进行 Friedman 检验的统计值和 p 值，以及在 $\alpha=0.01$ 水平下的显著性判别。

表 7 不同的 ε 下 RCP_Multiple_Opt 等 5 种算法的检验结果

Table 7 Test results of 5 algorithms such as RCP_Multiple_Opt with different ε

ε	χ^2_F	p-value	significance
0.1	54.587	3.966×10^{-11}	True
0.2	61.120	1.687×10^{-12}	True
0.5	55.013	3.228×10^{-11}	True
0.8	45.723	2.813×10^{-9}	True
1.0	41.920	1.733×10^{-8}	True
2.0	25.280	4.419×10^{-5}	True
5.0	14.480	0.005911	True
8.0	14.853	0.005015	True
10.0	12.293	0.015300	False

从图 3 可以看出，随着 ε 的逐渐增加，χ^2_F 的趋势逐渐减缓，这表明算法间的性能差异逐渐减弱，即差分隐私噪声对算法精度的影响逐渐减弱，这是自然的。而 p 值也由最初很小的值上升到大于给定的显著性水平 α，因此认为在 $\alpha=0.01$ 水平下，这 5 种算法在 $\varepsilon=10$ 时性能差异已经不明显了。

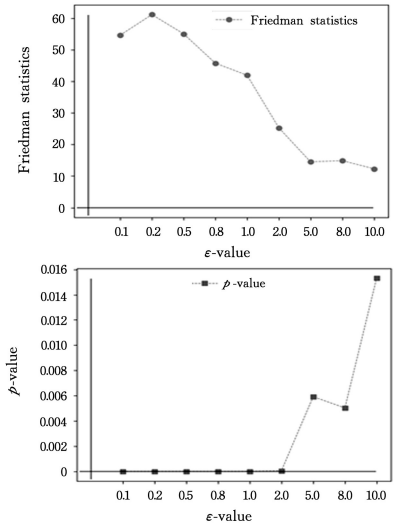


图 3 不同的 ε 下 Friedman 统计值和 p 值的比较

Fig. 3 Comparison of statistical value and p values of Friedman with different ε

请注意，一开始 p 值很小的原因是 RCP_Multiple_Opt 算法的平均排序值同其他算法相比小得多，从而拉开了差距。但是由于 RCP_Multiple，Duchi-EVP 和 Harmony 属于同类型算法，因此均秩的差异始终不会太大。

5.5.3 不同 ε 条件下基于 RMSE 的性能分析

表 5 列出了 $\varepsilon=0.1$ 时的 RMSE 均值，接下来求出 $\varepsilon>0.1$ 时 5 种算法的 RMSE 均值，如表 8 所列。

表 8 不同的 ε 下 RCP_Multiple_Opt 等 5 种算法的 RMSE 均值

Table 8 RMSE mean values of 5 algorithms such as RCP_Multiple_Opt with different ε

ε	RCP_Multiple_Opt	RCP_Multiple	Duchi-EVP	PM	Harmony
0.1	0.060946	0.087023	0.084699	0.071142	0.081232
0.2	0.042970	0.053252	0.052240	0.045901	0.050133
0.5	0.014739	0.019028	0.021552	0.017623	0.018670
0.8	0.010299	0.012596	0.012454	0.011414	0.012317
1.0	0.008400	0.011980	0.011128	0.009587	0.009924
2.0	0.005041	0.007023	0.006709	0.006432	0.005873
5.0	0.004217	0.005981	0.005836	0.004720	0.005370
8.0	0.004369	0.004622	0.004689	0.003905	0.004485
10.0	0.004179	0.004739	0.004892	0.004245	0.004501

同时，我们按照定理 3，设计“纯”RC 协议的多属性算法 No-DP。No-DP 仅包含 RC 噪声扰动，并不包含差分隐私噪声部分，No-DP 的估计方差理论上是所有 RCP 机制的估计方差的下界。

将 No-DP 分别在测试数据集上运行，并计算 RMSE 均值，结果为 0.004289，将该值同表 8 中 5 种算法的 RMSE 均值进行比较，结果如图 4 所示。

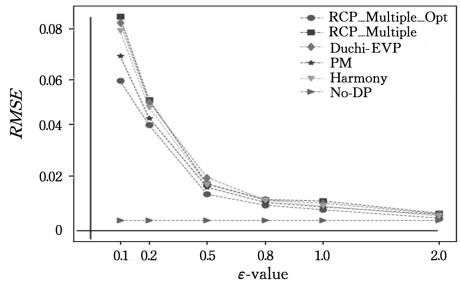


图 4 不同的 ε 下 RCP_Multiple_Opt 等 6 种算法 RMSE 均值的比较

Fig. 4 Comparison of RMSE mean values of 6 algorithms such as RCP_Multiple_Opt with different ε

从图 4 可以看出，所有算法的精度都随着 ε 的增大而提高。从指标上看，算法在 $\varepsilon=0.8$ 之前 RMSE 快速下降，在 ε 的值线性稳定增长（即隐私保证稳步下降）的前提下，性能提升迅速；当 $\varepsilon>0.8$ 后，性能提升趋于平稳，这时再降低隐私保护的程度，也不能使算法效用得到大幅提升。因此作为隐私-效用权衡，关于 RMSE 指标的拐点区域[0.5，0.8]的认知是非常关键的。

我们还可以看到，无论处于哪种隐私状态，RCP_Multiple_Opt 的平均 RMSE 变化趋势同其他算法一致，但其平均效用指标在绝大多数时候比其他方法小。亦即是说，在不同数据集上体现更好性能的概率较其他算法大。这是 Opt 算法使用效用优化定理在更优尺度区间执行的结果。该处理方式使得算法在满足 ε-LDP 条件的同时获得更好的统计精度。

此外，我们看到所有算法的 RMSE 均值都随着 ε 的增加逐渐逼近 No-DP 的 0.004289。通过进一步的实证研究可知，在 ε 足够大时，尽管由于噪声的随机性，偶尔会有算法的 RMSE 均值低于 0.004289，但所有算法的 RMSE 均值收敛于 0.004289 附近这一趋势保持不变，这进一步验证了定理 3 的结论。

5.5.4 通信开销

本实验提及的通信开销特指在上述给定样本量 $n=800000$、维数 $m=36$，并在 1—30 号待测试数据集上运行所需通信数据总量和通信总时耗。

其中，通信数据总量可以由用户端和服务器端之间传输的“单次报告比特数”乘以“通信轮次”来度量。由于上述 5 种算法均采用公币随机性提高通信效率，因此通信轮次均为 $O\left(\frac{n}{m}\right)$。在随机化器 $\mathcal{R}_0$ 是公币时，RCP_Multiple_Opt 和 RCP_Multiple 只需要向服务器端发送 0 或 1，因此单次开销刚好是 1 位。而 Duchi-EVP 需要向服务器端发送 $\frac{e^\varepsilon+1}{e^\varepsilon-1}$

或$-\frac{e^{\varepsilon}+1}{e^{\varepsilon}-1}$，Harmony需要发送$\frac{e^{\varepsilon}+1}{e^{\varepsilon}-1}m$或$-\frac{e^{\varepsilon}+1}{e^{\varepsilon}-1}m$，从理论上看开销较大，但由于$\varepsilon$和$m$是公币，因此用户端只需要向服务器端的聚合器发送1或-1，聚合器利用参数ε和m计算出相应的用户端报告值即可。因此，Duchi-EVP与Harmony的单次开销时间与RCP_Multiple_Opt和RCP_Multiple一样是1位。

根据PM的定义，若用户端数据为x_i，则以高概率输出$\omega_i \in [l(x_i), r(x_i)] \subset [-s, s]$，低概率输出其他值，因此PM需要发送浮点数，单次开销较大。

因此，我们从通信数据总量上看，算法RCP_Multiple_Opt，RCP_Multiple，Duchi-EVP和Harmony相同，均为$O\left(\frac{n}{m}\right)$；而PM较大，具体差异程度视PM发送的浮点数类型而定。

由于本实验针对的是电量有限、通信带宽受限的移动设备上的应用，因此我们将通信总时耗定义为数据在算法管线$f: \mathscr{D} \to \mathbb{R}$上通过的时长$T_f$。设用户端的随机化器$\mathscr{M}_i$的数据处理时间为$T_{\mathscr{M}}$，用户端的随机化器$\mathscr{M}_i$到服务器端的聚合器$\mathscr{Z}$的通信时间为$T_{MZ}$，服务器端的聚合器$\mathscr{Z}$的数据处理时间为$T_Z$，则：

$$T_f = T_M + T_{MZ} + T_Z$$

请注意，T_f忽略了多属性态下的均匀随机化器$\mathscr{R}_0$。这是由于5种算法的用户端使用相同的$\mathscr{R}_0$，在m相同时运行时耗也相同。

易知，RCP_Multiple_Opt，RCP_Multiple，Duchi-EVP和Harmony用户端均基于Bernoulli随机化器$B(1, p)$做一次抽样，分布参数p各不相同，但这一点并不会对抽样时间造成影响。因此这4种算法的T_M一致。而PM的用户端需要计算子区间$[l(x_i), r(x_i)]$，且报告值由两次均匀抽样获取，因此T_M较大。

$\mathscr{M}_i$到$\mathscr{Z}$的通信时间主要与单轮通信比特数、通信轮次、通信频率、设备带宽和其他硬件性能相关，由之前的通信数据总量分析可知：在固定软件、硬件与网络的环境下，PM的T_{MZ}值较其他算法更大。

由于实验中的5种算法在服务器端均采用累加器+线性变换的方式聚合，因此服务器端的运行时间T_Z是一致的。

由上述分析可知，从通信的总时耗来看，RCP_Multiple_Opt，RCP_Multiple，Duchi-EVP和Harmony通信开销基本相同，而PM开销较大，通信效率较低。

下面通过实验比较5种算法通信总时耗是否有显著性差异。

很容易看出，隐私预算ε仅与随机化器$\mathscr{M}_i$和聚合器$\mathscr{Z}$的计算式有关，这一点并不会影响计算效率，因此ε与通信开销无关。我们不妨固定$\varepsilon=1$，在1－30号数据集上分别运行5种算法，得到各种算法在各个数据集上的通信总时耗和平均总时耗，如表9所列。从表9可以看出，RCP_Multiple_Opt与RCP_Multiple，Duchi-EVP和Harmony在平均总时耗上的差异并不特别明显，而PM时耗较长，这与我们前面的分析是一致的。

表9 $\varepsilon=1$时RCP_Multiple_Opt等5种算法的通信总时耗

Table 9 Total communication time of 5 algorithms such as RCP_Multiple_Opt when $\varepsilon=1$

（单位：s）

Dataset Index	RCP_Multiple_Opt	RCP_Multiple	Duchi-EVP	PM	Harmony
1	76.543144	73.130275	76.186528	74.259419	80.867125
2	74.214661	71.390276	78.478882	86.014408	70.316290
3	78.458522	73.068251	70.455298	80.765612	69.139514
⋮	⋮	⋮	⋮	⋮	⋮
28	73.509330	81.281316	70.306858	81.153079	81.922351
29	73.724164	77.974089	72.643681	85.365669	76.366888
30	76.951257	80.962113	71.311125	83.723635	70.245282
Average total time	74.959208	76.843696	77.209752	80.825252	75.109780

图5中，RCP_Multiple_Opt，Duchi-EVP与Harmony时耗的中位数较小，RCP_Multiple的中位数其次，PM的中位数最大，这与表9得出的结论相似。此外，尽管图中RCP_Multiple_Opt，RCP_Multiple，Duchi-EVP和Harmony的中位数相差不大，但集中趋势差异明显。其中，RCP_Multiple_Opt的极差和四分位差最小，即数据最集中，而RCP_Multiple，Duchi-EVP和Harmony数据的离散程度较高。这说明通过算法RCP_Multiple_Opt获取的结果稳定，接近真实通信时耗的概率更大，数据代表性更好，而其他算法的代表性较差。

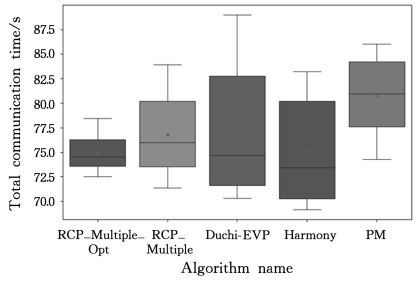


图5 $\varepsilon=1$时RCP_Multiple_Opt等5种算法通信总时耗的箱线图

Fig. 5 Boxplot of total communication time of 5 algorithms such as RCP_Multiple_Opt when $\varepsilon=1$

我们对表9中的数据进行方差分析，得到的结果如表10所列。

表10 RCP_Multiple_Opt等5种算法的通信总时耗数据的方差分析

Table 10 Variance analysis of total communication time data of 5 algorithms such as RCP_Multiple_Opt

	Sum_sq	Mean_sq	F-value	p-vlaue
Algorithm	224.381995	56.095499	2.383314	0.065394
Residual	1059.154270	23.536762		

表10中，p值0.065394＞0.05并不足以拒绝原假设。

亦即是说，分析结果无法否定上述算法的通信开销相近的论断。这表明，尽管能从数据上直观地看出 5 种算法在通信开销上有差距，RCP_Multiple_Opt，Duchi-EVP 与 Harmony 性能较好，PM 性能较差，但这个差距是有限的，至少在 $\alpha=0.05$ 的显著性水平下，没有足够证据支持 5 种算法在通信开销上有明显差异。

综合上述分析我们能了解到：从通信开销来看，RCP_Multiple_Opt，RCP_Multiple，Duchi-EVP 和 Harmony 的性能指标接近，比 PM 稍好。但从总体上看，5 种算法的差异并不特别明显。

5.6 结论

总的来说，所有的实证研究表明：优化算法 RCP_Multiple_Opt 在处理高隐私态数据方面具有相当大的性能优势，主要的原因是采用了 RCP 机制的效用优化定理，有效减小了估计误差，在提高算法效用的同时保证了差分隐私。

此外，在低隐私状态下，尽管所有算法的方差界限较为接近，但一方面由于推论 2 的保证，使得 RCP_Multiple_Opt 可以取得最优方差界限 $\mathcal{V}_{\text{opt}}^{(j)}$，这让 RCP_Multiple_Opt 在 ε 较大的情况下仍具有相对较高的精度；另一方面，RCP_Multiple_Opt 构造精简，仅传输 0，1 比特位，使得 RCP_Multiple_Opt 传输效率较高，传输时延较稳定。因此，Opt 算法的综合性能优于现有的同类型算法。

进一步来说，本实验也通过实证验证了定理 1—定理 4 和推论 1、推论 2 的正确性，实际展现了如何基于该理论来构造出更加简洁和高效的 LDP 均值估计算法。

结束语 本地差分隐私是一个相对较新的领域，还有很长的路要走。我们的研究也刚刚起步，很多问题仍然悬而未决。就 RCP 机制而言，均值估计的极值点产生于非对称区间是一个反直觉的结论，如何运用效用优化理论对尺度区间进行优化仍有很多工作要做。例如，我们现在采用的方式是以压缩尺度区间为主，而扩张该区间是否能够获取到同样甚至更好的效果，是一个有趣的问题。

我们希望本文能够起到弥合本地差分隐私体系中理论与实践之间鸿沟的作用，我们相信在这个领域还有很多未知的潜力，在探索和挖掘过程中恐怕还要迎接更多挑战，希望我们的研究能为此尽一些绵薄之力[28-32]。

参 考 文 献

[1] DWORK C. Differential Privacy[C]// The 33rd International Colloquium on Automata, Languages and Programming (ICALP). 2006: 1-12.

[2] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[M]// Theory of Cryptography. Springer, 2006.

[3] DWORK C, ROTH A. The algorithmic foundations of differential privacy[J]. Foundations and Trends in Theoretical Computer Science, 2014, 9(3/4): 211-407.

[4] DWORK C, LEI J. Differential Privacy and Robust Statistics [C]// Proceedings of the 41st ACM Symposium on Theory of Computing. New York: Association for Computing Machinery, 2009: 371-380.

[5] KASIVISWANATHAN S, LEE H K, NISSIM K, et al. What can we learn privately? [J]. SIAM Journal on Computing, 2011, 40(3): 793-826.

[6] WANG T, ZHANG X F, FENG J Y, et al. A Comprehensive Survey on Local Differential Privacy Toward Data Statistics and Analysis[J]. Sensors, 2020, 20(24): 7030.

[7] YANG M M, LYU L J, ZHAO J, et al. Local Differential Privacy and Its Applications: A Comprehensive Survey[J]. arXiv: 2008.03686, 2020.

[8] WARNER S L. Randomized response: A survey technique for eliminating evasive answer bias[J]. Journal of the American Statistical Association, 1965, 60(309): 63-69.

[9] DUCHI J, JORDAN M, WAINWRIGHT M. Minimax optimal procedures for locally private estimation[J]. Journal of the American Statistical Association, 2018, 113(521): 182-201.

[10] DUCHI J, JORDAN M, WAINWRIGHT M. Privacy aware learning[J]. Journal of the ACM(JACM), 2014, 61(6): 38.

[11] DUCHI J, JORDAN M, WAINWRIGHT M. Local privacy and statistical minimax rates[C]// 54th Annual Symposium on Foundations of Computer Science. IEEE, 2013: 429-438.

[12] KAIROUZ P, OH S, VISWANATH P. Extremal mechanisms for local differential privacy[J]. Advances in Neural Information Processing Systems, 2014, 2(20): 2879-2887.

[13] WANG N, XIAO X K, YANG Y, et al. Collecting and analyzing multidimensional data with local differential privacy[C]// 35th International Conference on Data Engineering(ICDE). IEEE, 2019: 638-649.

[14] LI Z T, WANG T H, LOPUHAA-ZWAKENBERG M, et al. Estimating Numerical Distributions under Local Differential Privacy[J]. arXiv: 1912.01051, 2019.

[15] NGUYEN T, XIAO X K, YANG Y, et al. Collecting and analyzing data from smart device users with local differential privacy [J]. arXiv: 1606.05053, 2016.

[16] DALENIUS T, VITALE R A. A new randomized response design for estimating the mean of a distribution[M]// Contributions to Statistics. 1979: 43-47.

[17] GREENBERG B G, KUEBLER J R R R, ABERNATHY J R, et al. Application of the randomized response technique in obtaining quantitative data[J]. Journal of the American Statistical Association, 1971, 66(334): 243-250.

[18] CHAUDHURI A, MUKHERJEE R. Randomized response techniques: a review[J]. Statistica Neerlandica, 1987, 41(1): 27-44.

[19] WARNER S L. The linear randomized response model[J]. Journal of the American Statistical Association, 1971, 66(336): 884-888.

[20] WARNER S L. Optimal randomized response models[J]. International Statistical Review/Revue Internationale de Statistique, 1976, 15(8): 205-212.

[21] BASSILY R, SMITH A. Local, Private, Efficient Protocols for Succinct Histograms[C]// Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing. ACM, 2015: 127-135.

[22] ACHARYA J, SUN Z, ZHANG H. Hadamard response: Esti-

mating distributions privately, efficiently, and with little communication[C]// Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics(AISTATS). 2019.

[23] WANG T H, LOPUHAA-ZWAKENBERG M, LI Z T, et al. Locally Differentially Private Frequency Estimation with Consistency[C]// 27th Annual Network and Distributed System Security Symposium, NDSS 2020. San Diego, California, USA, 2020.

[24] ERLINGSSON U, PIHUR V, KOROLOVA A. Rappor: Randomized aggregatable privacy-preserving ordinal response[C]// Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2014: 1054-1067.

[25] BEBENSEE B. Local differential privacy: a tutorial[J]. arXiv: 1907.11908, 2019.

[26] BASSILY R. Linear queries estimation with local differential privacy[C]// AISTATS. 2019.

[27] WANG T, BLOCKI J, LI N, et al. Locally differentially private protocols for frequency estimation[C]// USENIX Security. 2017.

[28] ZHANG X J, FU N, MENG X F. Towards Spatial Range Queries Under Local Differential Privacy[J]. Journal of Computer Research and Development, 2020, 57(4): 847-858.

[29] ZHANG X J, FU N, MENG X F. Key-Value Data Accurate Collection under Local Differential Privacy[J]. Chinese Journal of Computers, 2020, 43(8): 1479-1492.

[30] YE Q Q, MENG X F, ZHU M J, et al. Survey on local differential privacy[J]. Ruan Jian Xue Bao/Journal of Software, 2018, 29(7): 1981-2005.

[31] ZHU S X, WANG L, SUN G L. A Perturbation Mechanism for Classified Transformation Satisfying Local Differential Privacy [J]. Journal of Computer Research and Development, 2021, 59(2): 430-439.

[32] LIU Y X, CHEN H, LIU Y H, et al. Privacy-preserving Techniques in Federated Learning[J]. Ruan Jian Xue Bao/Journal of Software, 2022, 33(3): 1057-1092.

LIU Likang, born in 1977, master, senior engineer, is a member of China Computer Federation. His main research interests include statistical privacy, dimensionality reduction and analysis of ultra-high-dimensional data streams, machine learning and optimization theory.

ZHOU Chunlai, born in 1976, Ph.D, associate professor. His main interests interests include trustworhty AI and privacy computing.

(责任编辑:喻藜)