软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2017,28(9):2309−2322 [doi: 10.13328/j.cnki.jos.005184]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

线性查询的一种近似最优差分隐私机制*

武跟强[1,2], 贺也平[1,3], 夏娴瑶[1]

[1](中国科学院 软件研究所 基础软件国家工程研究中心,北京 100190)

[2](兰州财经大学 信息工程学院,甘肃 兰州 730020)

[3](计算机科学国家重点实验室(中国科学院 软件研究所),北京 100190)

通讯作者: 武跟强, E-mail: genqiang80@gmail.com; 贺也平, E-mail: yeping@nfs.iscas.ac.cn

摘 要: 在差分隐私保护程度确定的条件下使数据的有用性最大化的问题,称为差分隐私的最优机制问题.最优机制问题是差分隐私理论中的一个重要问题,与差分隐私模型的理论基础及应用前景有直接联系.与已有的研究不同,提出一种不基于敏感度的分析方法来寻找最优机制:首先,将最优机制问题构造为一个多目标函数优化问题,并提出了一种差分隐私机制构造方法,在此基础上,对线性查询问题给出了一种近似最优差分隐私机制,该机制达到了差分隐私不等式的边界.此外,大部分分析方法也可对非线性查询的最优机制问题进行分析.该研究揭示了敏感度方法的不足之处,发现其无法刻画数据集的邻居集合对应的查询函数值集合的特性,而该集合包含了差分隐私的一些深层特征.

关键词: 线性查询;差分隐私;最优机制;多目标优化;非敏感度方法

中图法分类号: TP309

中文引用格式: 武跟强,贺也平,夏娴瑶.线性查询的一种近似最优差分隐私机制.软件学报,2017,28(9):2309−2322. http://www.jos.org.cn/1000-9825/5184.htm

英文引用格式: Wu GQ, He YP, Xia XY. Near-Optimal differentially private mechanism for linear queries. Ruan Jian Xue Bao/ Journal of Software, 2017,28(9):2309−2322 (in Chinese). http://www.jos.org.cn/1000-9825/5184.htm

Near-Optimal Differentially Private Mechanism for Linear Queries

WU Gen-Qiang[1,2], HE Ye-Ping[1,3], XIA Xian-Yao[1]

[1](National Engineering Research Center of Fundamental Software, Institute of Software, The Chinese Academy of Sciences, Beijing 100190, China)

[2](School of Information Engineering, Lanzhou University of Finance and Economics, Lanzhou 730020, China)

[3](State Key Laboratory of Computer Science, Institute of Software, The Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The optimal differentially private mechanism problem is to maximize the data utility on a fixed privacy protection extent. The optimal mechanism problem is an important topic in differential privacy, which has close connection with both theoretical foundation and future applications of differential privacy model. This paper proposes a analyzing method about the topic, which is not based on the sensitivity method. First, the optimal mechanism problem is constructed to be a multi-objective optimization problem, and a new method for constructing differentially private mechanism is introduced. Then, a near-optimal mechanism is provided for the linear queries, which reaches the boundary of the differential privacy inequality. Although this paper focuses on the linear queries, most part of the analyzing method introduced is applicable to the non-linear queries. This paper finds the drawback of the sensitivity method and uncovers some deeper characteristics of differential privacy.

Key words: linear query; differential privacy; optimal mechanism; multi-objective optimization; non-sensitivity method

* 基金项目: 中国科学院战略性先导科技专项基金(XDA06010600)
Foundation item: Strategic Priority Research Program of the Chinese Academy of Sciences (XDA06010600)
收稿时间: 2016-07-10; 修改时间: 2016-09-04; 采用时间: 2016-11-10; jos 在线出版时间: 2017-02-20
CNKI 网络优先出版: 2017-02-20 14:02:21, http://www.cnki.net/kcms/detail/11.2560.TP.20170220.1402.012.html

1 引 言

大数据中的信息挖掘,是当今计算机科学领域中的一个热门研究领域.然而,绝大多数大数据中都包含了很多敏感的个人隐私信息,如医疗记录、搜索引擎记录、电子商务数据、社会网络数据等[1,2].如何以隐私保持的方式在隐私敏感数据集中挖掘有用信息,在近二、三十年中一直是一个很活跃的研究领域[3].特别地,随着差分隐私模型的引入,隐私保持信息分析和挖掘在近年中得到了迅速的发展[4−8].然而,这个领域还处于比较初级的发展阶段,很多关键问题还没能得到解决.尽管差分隐私是一个有严格数学定义的隐私模型,通过其可对数据集中的个人隐私信息进行很好的保护,但是,差分隐私模型也因其较差的数据有用性而得到很多批评——经差分隐私处理后数据的有用性要比经类 k-匿名(如 k-匿名、ℓ-多样性、t-邻近性等)模型处理后数据的有用性差[9,10](当然,类 k-匿名模型的隐私保护程度差也是其重大缺陷).然而,经差分隐私处理后数据的有用性较弱的原因一直没能被充分研究.也就是说,没有比较完整的结论,说明经差分隐私处理后数据有用性差的问题是因为差分隐私模型的特性,还是因为现在的研究还没能找到更好的相关差分隐私算法,或者是因为隐私保护本身的特性而与隐私模型选择无关.关于这个问题,学者们从不同的角度(直接或间接地)进行了研究,如统计学习角度[11−13]、数据类型角度[14,15]、数据分析角度[16]、最优机制角度[17−19]等.本文的工作属于最优机制研究,即:对一个特定的查询问题,我们试图在保持差分隐私的条件下达到最优的数据有用性.本文构造了一些基本的分析方法来对差分隐私的最优机制问题进行分析.

对多查询问题(即同时对多个查询进行处理),从查询函数的角度,(广义)最优机制问题可分为两个方面:函数压缩和(狭义)最优机制.函数压缩关注不同查询之间的相关性问题,如线性相关性.函数压缩试图将相关性很高的多查询问题压缩为相关性较低的少量查询问题,然后通过寻找后者的差分隐私机制来求解前者的差分隐私机制,由此来减小噪声复杂度[20−22].(狭义)最优机制关注于低相关性函数查询的最优机制寻找问题.函数压缩问题可表述如下:对于相关性较高的 m 个查询问题 $g_1(x),\ldots,g_m(x)$,经过某种压缩算法,可将其压缩为 $n(<m)$个相关性很低的查询函数 $f_1(x),\ldots,f_n(x)$.最简单情形,存在一个 $m\times n$ 实值矩阵 A,满足 $g(x)=Af(x)$,其中,$g(x)=(g_1(x),\ldots,g_m(x))^T$,$f(x)=(f_1(x),\ldots,f_n(x))^T$.若矩阵 A 是一个小元素矩阵(即 A 的每个元素都很小),则通过研究关于查询函数 $f(x)$ 的差分隐私机制,就可以以后处理(post-processing)原理[5]通过表达式 $g(x)=Af(x)$来找到 $g(x)$的差分隐私机制.与文献[18]类似,本文主要讨论(狭义)最优机制问题.也就是说:对于相关性很低的函数 $f(x)$,需要找到最优差分隐私机制,从而得到数据有用性最高的(或最精确的)对 $f(x)$的查询反馈.在本文以下的分析中,我们都假定多查询函数是低相关的,因此无需关注其函数压缩.

1.1 本文贡献

本文试图构造一些基本的分析方法来对差分隐私的最优机制问题进行分析.

首先,不同于文献[18–24]中将最优机制问题归结为一个单目标函数优化问题,本文引入了多目标函数优化模型来求解最优机制.这种多目标函数模型与其他单目标函数模型相比,如 Min-Max 目标函数模型[18,19],具有很大优点:多目标模型将不同数据集对应分布的有用性的权衡问题剥离出了最优差分隐私机制问题,这为分析差分隐私机制的最优性提供了方便.

其次,本文引入了一种新的最优差分隐私机制分析方法.该分析方法受到了文献[25]中的梯子函数(ladder function)的启发,将概率分布通过集合切分及合理测度赋值来达到数据有用性和隐私性的权衡.

最后,针对线性查询问题,本文寻找到了线性查询问题的一种近似最优机制(定理 2).据我们所知,定理 2 是关于线性查询最优机制的最全面的结论.并且,本文的大部分分析方法对分析非线性查询问题的最优机制也是适用的,比如,第 3 节、第 4 节的内容都没有对函数类型进行限定.本文的深层目的是探讨差分隐私的深层数学结构.本文发现:一个数据集的邻居集合对应的查询函数值集合里面包含了差分隐私的一些非常重要的特征,而敏感度只是该集合的极值特性.本文通过研究该集合中点的分布情况来探讨差分隐私的深层特征,相关内容请参考第 6 节的讨论.

1.2 相关工作

差分隐私的研究主要从 3 个关注点展开:数据隐私性、数据有用性(或噪声复杂度)及算法复杂度.对一个查询问题,这三者之间相互冲突,需要在他们之间进行权衡.

本文的研究关注于数据隐私性及数据有用性之间的权衡问题,不考虑相关算法的复杂度问题.即:考虑在隐私保护程度确定的条件下,获得有用性最高的查询结果时的机制寻找问题(即最优机制问题).前人分别从不同的角度对一些特殊的线性查询的最优机制问题进行了研究,如单输出值查询[18,19,24,29]、多输出连续实值查询[23]、多输出值查询噪声复杂度上下界及其逼近[17,28]等.不同于先前的工作,本文对一般的线性查询问题寻找最优机制,并得出了相关结果(定理 2),扩展了在线性查询最优机制方面的结论.本文的分析方法其实是为了分析非线性查询问题而发展起来的一套分析理论,因此也可以对非线性查询问题进行分析.

如第 1 节的分析,另一类研究关注于多线性查询问题中各查询之间的相关性(如线性相关性)问题,并试图将相关性很高的多查询问题压缩为相关性很低的少量查询以降低噪声复杂度[20−22].这一类研究关注于线性查询相关算法的隐私性、数据有用性及算法复杂度这 3 方面的权衡,但并不强调相关机制的(噪声复杂度)最优性.

从统计学习的角度,很多学者试图将关于数据集 x 的多线性查询转化为一个关于数据集 y 的对应多线性查询[30−32].该方法的理论基础是基于查询函数的统计特性及自变量的数据集的特性.这类研究关注噪声复杂度、计算复杂度的估计及与隐私之间的权衡,没有强调最优机制问题.

敏感度方法是差分隐私中的一类重要方法,如全局敏感度[27]、平滑敏感度[26]及局部敏感度[25]等.现在的绝大多数差分隐私算法都是通过敏感度方法来进行隐私分析的.本文的分析方法受到了文献[25]中梯子函数(ladder function)的极大启发.但文献[25]中梯子函数的方法还是基于(局部)敏感度,而本文将该方法进行了改进,使其脱离了敏感度方法,并发展出一套与文献[25]完全不同的分析方法.并且,文献[25]并不讨论机制的最优性及方法的推广.关于本文方法与敏感度方法之间的优劣比较见本文第 6 节的讨论.据我们所知,本文方法是第一个不基于敏感度而能对非线性查询最优机制进行分析的方法.

2 问题形式化

在本文中,数据集(dataset)是一些记录(行向量)的多重集(multiset),每个记录代表个体信息,其形式化表示为[5,17,21]:设$\mathcal{X}$是所有不同记录的集合,且设一个数据集 x 是$\mathcal{X}$中一些记录的(多重)集合.本文将数据集 x 表示为直方图 $x\in\mathbb{N}^{|\mathcal{X}|}$,其中,每个分量 x_i 表示数据集 x 中类型 $i\in\mathcal{X}$的个数($\mathbb{N}$表示包含 0 的自然数集,$|\mathcal{X}|$表示$\mathcal{X}$中元素的(可以不可数)个数).两个数据集 x,y 的距离定义为他们直方图表示的差值的ℓ_1-范数,即$\|x-y\|_1$.若$\|x-y\|_1=1$,则称 x,y 为邻居或相邻数据集.数据集 x 的所有邻居的集合记为$\mathcal{N}^x$.设 $\mathcal{D}\subseteq\mathbb{N}^{|\mathcal{X}|}$ 是所有可允许的数据集的集合,则$\mathcal{N}^x=\{x'\in\mathcal{D}:\|x-x'\|_1=1\}$.

本文主要研究线性查询(linear query)的最优机制.线性查询是一类很广泛的查询函数[5,17,21],包括计数函数、列联表(contingency table)及范围查询(range query)等.线性查询 f 定义为 $f:\mathcal{X}\to\mathbb{R}$,且对有限元素数据集 x(当$\mathcal{X}$为离散集合时):

$$f(x)=\sum_{i=1}^{|\mathcal{X}|}x_i f(\mathcal{X}_i),$$

或(当$\mathcal{X}$为连续集合时):

$$f(x)=\sum_{i=1}^{|\mathcal{Y}|}y_i f(\mathcal{Y}_i),$$

其中,$\mathcal{Y}$表示有限元素数据集 x 中非零元的不同类型的集合,y_i 是类型$\mathcal{Y}_i$ 的个数.我们称函数 f 关于数据集 x 的邻居集为$\{f(x'):x'\in\mathcal{N}^x\}$,记为$\mathcal{V}^x$.显然,若 f 为线性函数,则$\mathcal{V}^x=f(x)\pm\{f(y):y\in\mathcal{X}\}$,因为$\{f(y):y\in\mathcal{X}\}$只与 f 有关,我们称其为线性函数 f 的邻居集,记为$\mathcal{V}$.集合$\mathcal{V}$中包含了很多重要的差分隐私所需的特征,如(全局或局部)敏感度 $\Delta f=\max_{a,b\in\mathcal{V}}|a-b|$.本文试图通过集合$\mathcal{V}$的其他特性(集合中点的分布)来理解差分隐私的一些深层数学结构.

n 个线性查询 $f_1(x),\dots,f_n(x)$可定义为一个多输出值线性函数 f 的查询问题.当$\mathcal{X}$为离散集合时,定义:

$$f(x)=Ax,$$

其中,矩阵 A 的元素 $a_{ij}=f_i(\mathcal{X}_j)$,数据集 x 为其直方图表示的列向量.当$\mathcal{X}$为连续集合时,定义:

$$f(x)=Ay,$$

其中,y 为数据集 x(有限集)的直方图表示中非零元组成的列向量$(y_1,\dots,y_k)^T$,矩阵 A 的元素 $a_{ij}=f_i(\mathcal{Y}_j)$,$\mathcal{Y}_j$ 代表非零元 y_j 对应的记录类型.因此,一个(多)线性查询函数可统一记为 $f:\mathcal{X}\to\mathbb{R}^n$;或为了表述的方便,记为 $f:\mathcal{D}\to\mathbb{R}^n$,其中,$\mathcal{D}$ 为所有可允许的取值于$\mathcal{X}$的数据集的集合.

差分隐私机制[5]是一类随机函数,其输出是一个随机变量.设$\mathcal{D}$为所有可允许的数据集的集合,且对查询函数 f,设$\mathcal{R}=\{f(x):x\in\mathcal{D}\}$.

定义 **1**. 以$\mathcal{D}$为定义域,以$\mathcal{R}$为值域的随机函数$\mathcal{M}$关联了一个确定性函数$M:\mathcal{D}\to\Delta(\mathcal{R})$,其中,$\Delta(\mathcal{R})$代表所有$\mathcal{R}$上的概率分布的集合.当输入为 $x\in\mathcal{D}$时,$\mathcal{M}$输出随机变量$\mathcal{M}(x)$,其服从概率分布 $M(x)$.$\mathcal{M}$满足ε-差分隐私,如果对任意的相邻数据集 x,y 及对任意的可测集 $S\subseteq\mathcal{R}$,满足:

$$\Pr[\mathcal{M}(x)\in S]\le\exp(\varepsilon)\times\Pr[\mathcal{M}(y)\in S].$$

定义 1 说明,差分隐私机制$\mathcal{M}$给每一个数据集 x 对应了一个概率分布 $M(x)$,而相邻数据集所对应的概率分布需满足定义 1 中的不等式.因此,差分隐私机制$\mathcal{M}$与概率分布族$\{M(x):x\in\mathcal{D}\}$是一一对应的.若$\{M(x):x\in\mathcal{D}\}$是连续概率分布族,其密度函数族为$\{p_x:x\in\mathcal{D}\}$,则差分隐私机制$\mathcal{M}$与密度函数族$\{p_x:x\in\mathcal{D}\}$一一对应.在以下的分析中,如果没有特殊说明,本文将$\mathcal{M}$与其密度函数族$\{p_x:x\in\mathcal{D}\}$视为相同的量.

当$\mathcal{R}=\mathbb{R}^n$时,我们定义函数 $f:\mathcal{D}\to\mathbb{R}^n$ 关于 x 的局部ℓ_p敏感度为 $\max_{x'\in\mathcal{N}^x}\|f(x)-f(x')\|_p$,定义 f 的全局ℓ_p敏感度为 $\max_{x\in\mathcal{D}}\max_{x'\in\mathcal{N}^x}\|f(x)-f(x')\|_p$,其中,$\|\cdot\|_p$ 代表 p-范数.

2.1 效用模型

本文以连续情形为例进行分析,离散情形可类似处理.对于线性查询函数 $f:\mathcal{D}\to\mathbb{R}^n$ 及数据集 x,根据定义 1,不同的差分隐私机制对应了不同的$\mathbb{R}^n$ 上的密度函数族,也就对应了不同的查询反馈值.本文研究满足ε-差分隐私的所有机制中查询反馈最精确的那个密度函数族.下面定义效用模型,用其来衡量差分隐私输出的有用性.

在差分隐私的研究中,有几种常用的衡量差分隐私输出有用性的模型,最常用的为 Min-Max 模型[17,18,23],即:

$$\arg\min_{\mathcal{M}}\sup_{x\in\mathcal{D}}\int_{r\in\mathbb{R}^n}\|f(x)-r\|_p p^x(r)\mathrm{d}r.$$

不同于 Min-Max 等单目标函数优化模型,本文使用多目标函数优化模型,即:

$$\arg\min_{\mathcal{M}}\{\int_{r\in\mathbb{R}^n}\|f(x)-r\|_p p^x(r)\mathrm{d}r:x\in\mathcal{D}\},$$

其中,$p^x(r)$代表$\mathcal{M}(x)$的密度函数,$\|\cdot\|_p$ 代表 p-范数(p 是一个常量).多目标函数模型意在寻找对所有数据集 x 都达到最优的差分隐私机制.而 Min-Max 等单目标函数模型需要对不同数据集对应输出的有用性进行权衡,这增加了问题分析的复杂度.也就是说,多目标模型将不同数据集对应输出的有用性权衡问题剥离出了优化机制问题,这为分析差分隐私机制的最优性提供了方便.

2.2 最优化问题

现在将最优差分隐私机制问题归纳为如下的多目标优化问题.

$$\left.\begin{aligned}&\arg\min_{p_x:x\in\mathcal{D}}\{\int_{r\in\mathbb{R}^n}\|f(x)-r\|_p p^x(r)\mathrm{d}r:x\in\mathcal{D}\}\\&\text{s.t. 对任意邻居}x,x'\text{及任意可测集}S\subseteq\mathbb{R}^n,\\&\quad\Pr[\mathcal{M}(x)\in S]\le\exp(\varepsilon)\times\Pr[\mathcal{M}(x')\in S]\end{aligned}\right\}\tag{1}$$

上面的优化问题等价于:

$$\left.\begin{array}{l}\underset{p_x:x\in\mathcal{D}}{\operatorname{argmin}}\{\int_{r\in\mathbb{R}^n}\| f(x)-r\|_p p^x(r)\mathrm{d}r : x\in\mathcal{D}\}\\ \text{s.t. 对任意邻居}x,x',\text{任意}r\in\mathbb{R}^n\setminus B,\mu(B)=0,\\ \quad p^x(r)\le e^{\varepsilon}p^{x'}(r)\end{array}\right\} \tag{2}$$

其中,μ代表$\mathbb{R}^n$上的 Lebesgue-Stieltjes 测度,$B\in\mathbb{R}^n$.为了表述的简洁性,在本文中设 $B=\varnothing$.p 是一个常量.

对多目标函数优化问题(1)或(2),需要寻找(可能多个)密度函数族$\{p_x:x\in\mathcal{D}\}$,其对优化问题(1)或(2)具有 Pareto 最优性.更多关于多目标函数优化问题的知识见文献[33].

3 差分隐私机制分解

对于集合$\mathcal{R}$上的密度函数 $p^x(r)$,可以将其表示为 $\exp(\varepsilon q(x,r))$的形式,其中, $q(x,r)=\frac{1}{\varepsilon}\ln p_x(r)$ 并设置 $\ln 0=-\infty$.因此,若函数 $q:\mathcal{D}\times\mathcal{R}\to\mathbb{R}_-$,其中,$\mathbb{R}_-$表示非正实数集与$-\infty$的并集,则对于数据集 x,$\exp(\varepsilon q(x,r))$可表示任何一个$\mathcal{R}$上的密度函数.因此,记密度函数 $p^x(r)=\exp(\varepsilon q(x,r))$,其中,$q:\mathcal{D}\times\mathcal{R}\to\mathbb{R}_-$称为 $p^x(r)$的质量函数.

对于数据集 x 及相应密度函数 $p^x(r)=\exp(\varepsilon q(x,r))$,下面根据质量函数 $q(x,r)$的值对集合$\mathcal{R}$进行切分.对数据集 x,设 $q(x,r)$的值域为$\mathcal{Q}^x$.对任意一个 $z\in\mathcal{Q}^x$,记 $\mathcal{R}_z^x=\{r\in\mathcal{R}:q(x,r)=z\}$.则 $\mathcal{R}=\bigcup_{z\in\mathcal{Q}}\mathcal{R}_z^x$且$\mathcal{R}_z^x\cap\mathcal{R}_{z'}^x=\varnothing$,其中 $z,z'\in\mathcal{Q}_x$ 且 $z\neq z'$.记 $\mathcal{R}^x=\{\mathcal{R}_z^x:z\in\mathcal{Q}^x\}$,则 $p^x(r)$与集合对$(\mathcal{Q}^x,\mathcal{R}^x)$一一对应.因此,差分隐私机制$\{p_x:x\in\mathcal{D}\}$与集合对族$\{(\mathcal{Q}^x,\mathcal{R}^x):x\in\mathcal{D}\}$形成一一对应.不同的集合对族$\{(\mathcal{Q}^x,\mathcal{R}^x):x\in\mathcal{D}\}$对应了不同的差分隐私机制.

对数据集 $x\in\mathcal{D}$,当$\mathcal{Q}^x$ 是一个可数点的集合时,记$\mathcal{Q}^x=\{z_0,z_1,\ldots\}$及 $\mathcal{R}^x=\{\mathcal{R}_0^x,\mathcal{R}_1^x,\ldots\}$,其中, $\mathcal{R}_i^x=\mathcal{R}_{z_i}^x$.则优化问题(2)等价于下面优化问题(3):

$$\left.\begin{array}{l}\underset{(\mathcal{Q}^x,\mathcal{R}^x):x\in\mathcal{D}}{\operatorname{argmin}}\left\{\sum_{i=0}^{\infty}\frac{\exp(\varepsilon z_i^x)}{\alpha_x}\int_{r\in\mathcal{R}_i^x}\| f(x)-r\|_p\,\mathrm{d}r : x\in\mathcal{D}\right\}\\ \text{s.t. 对任意邻居}x,x',\text{任意}i,j\in\mathbb{N}\text{且满足}\mathcal{R}_i^x\cap R_j^{x'}\neq\varnothing,\text{有}\\ \quad \exp(\varepsilon(z_i^x-z_j^{x'}-1))\le\frac{\alpha_x}{\alpha_{x'}}\end{array}\right\} \tag{3}$$

其中, $\alpha_x=\sum_{j=0}^{\infty}\exp(\varepsilon z_j^x)\int_{r\in\mathcal{R}_j^x}\mathrm{d}r$. 此时,记 $a_j=\int_{r\in\mathcal{R}_j^x}\| f(x)-r\|_p\,\mathrm{d}r, b_j=\int_{r\in\mathcal{R}_j^x}\mathrm{d}r$.

对数据集 $x\in\mathcal{D}$,当$\mathcal{Q}^x$ 是一个连续点集时,优化问题(2)等价于下面优化问题(4):

$$\left.\begin{array}{l}\underset{(\mathcal{Q}^x,\mathcal{R}^x):x\in\mathcal{D}}{\operatorname{argmin}}\left\{\int_{z\in\mathcal{Q}^x}\frac{\exp(\varepsilon z)}{\alpha_x}\sum_{r\in\mathcal{R}_z^x}\| f(x)-r\|_p\mathrm{d}z : x\in\mathcal{D}\right\}\\ \text{s.t. 对任意邻居}x,x',\text{任意}i,j\in\mathbb{N}\text{且满足}\mathcal{R}_i^x\cap R_j^{x'}\neq\varnothing,\text{有}\\ \quad \exp(\varepsilon(z_i^x-z_j^{x'}-1))\le\frac{\alpha_x}{\alpha_{x'}}\end{array}\right\} \tag{4}$$

其中, $\alpha_x=\int_{z\in\mathcal{Q}^x}\exp(\varepsilon z^x)\,|\,\mathcal{R}_z^x\,|\,\mathrm{d}z$.

4 极小差分隐私集序列

根据上一节的结论,差分隐私机制与集合对族$\{(\mathcal{Q}^x,\mathcal{R}^x):x\in\mathcal{D}\}$一一对应.因此,只需要研究集合对族的性质就可以研究清楚差分隐私机制的性质.本节讨论$\{\mathcal{R}^x:x\in\mathcal{D}\}$的构造问题,也就是说,研究满足差分隐私的对应机制中对值域$\mathcal{R}=\{f(x):x\in\mathcal{D}\}$的切分问题.首先给出一个假设,该假设说明:当查询值是 $f(x)$时,密度函数 $p^x(r)$应在 $r=f(x)$点达到最大.

假设 **1**. 对任意数据集 $x\in\mathcal{D}$及任意 $r\in\mathcal{R}$,有 $p^x(f(x))\ge p^x(r)$,且 $f(x)\in\mathcal{R}_0^x, z_0^x=0$.

注:假设 1 中, $z_0^x=0$ 可能导致整个值域$\mathcal{R}$上的概率和(或积分)不为 1,因此需要进行归一化处理.这在第 5 节有所体现.

基于假设 1,构造一类集合切分 $\mathcal{R}=\bigcup_t I_t^x, x\in\mathcal{D}$,这一类集合切分将相邻数据集的函数值尽可能放到同一个子集合或相邻的子集合,这样就便于分析和构造差分隐私机制.具体构造如下:对任一 $x\in\mathcal{D}$,设置 $I_0^x=\{f(x)\}$ 及 $I_t^x=\bigcup_{x'\in\mathcal{N}^x} I_{t-1}^{x'}\setminus\bigcup_{i=0}^{t-1} I_i^x, t>0$.这种集合切分构造借鉴了文献[25]中梯子函数的构造方法.该构造有如下性质.

引理 **1**. 对任意的邻居 $x,x'\in\mathcal{D}$,及对任意的$|t-s|>1$,有 $I_t^x\cap I_s^{x'}=\varnothing$且$\mathcal{R}=\bigcup_{i=0}^{\infty} I_i^x$.

证明:首先定义集合 $A_t^x=\{f(x'): x'\in\mathcal{D}, \|x-x'\|_1\le t\}$.下面证明 $A_t^x=\bigcup_{i=0}^{t} I_i^x$.首先,对每个 $i\le t$, I_i^x 是一些与 x 的距离小于等于 i 的数据集的函数值的集合,而 A_t^x 是所有与 x 的距离小于等于 t 的数据集的函数值的集合,因此 $\bigcup_{i=0}^{t} I_i^x\subseteq A_t^x$.下面使用数学归纳法来证明 $A_t^x\subseteq\bigcup_{i=0}^{t} I_i^x$.当 $t=0$ 时, $A_0^x=\bigcup_{i=0}^{t} I_i^x=\{f(x)\}$ 显然成立;假设对所有 $t<k$,有 $A_t^x\subseteq\bigcup_{i=0}^{t} I_i^x$;当 $t=k$ 时,对任意 $r\in A_k^x$,存在一个数据集 y 及整数 j,满足$\|x-y\|_1=j\le k$ 且 $f(y)=r$.显然, $r\in A_j^x$.若 $j<k$,根据假设,有 $r\in\bigcup_{i=0}^{j} I_i^x$.若 $j=k$,则存在 x',满足$\|x-x'\|_1=1$ 且$\|x'-y\|_1=k-1$.根据假设,有 $r\in A_{k-1}^{x'}\subseteq\bigcup_{i=0}^{k-1} I_i^{x'}$.再由 I_k^x 的定义,有 $r\in I_k^x$.因此, $A_k^x\subseteq\bigcup_{i=0}^{k} I_i^x$.从而,对任意 $t\ge 0, x\in\mathcal{D}$,有 $A_t^x=\bigcup_{i=0}^{t} I_i^x$.

由 $A_t^x=\bigcup_{i=0}^{t} I_i^x$,易得 $\mathcal{R}=\bigcup_{i=0}^{\infty} I_i^x$及$I_t^x=\bigcup_{x'\in\mathcal{N}^x} I_{t-1}^{x'}\setminus A_{t-1}^x$.下面证明 $I_t^x\cap I_s^{x'}=\varnothing$.因为 $I_t^x=\bigcup_{x'\in\mathcal{N}^x} I_{t-1}^{x'}\setminus A_{t-1}^x$, 则对任意 $r\in I_t^x$,存在 $y\in\mathcal{D}$,满足$\|x-y\|_1=t$ 及 $r=f(y)$,且对任何满足$\|x-y'\|_1\le t-1$ 的 $y'\in\mathcal{D}$,有 $r\ne f(y')$.另一方面,对任意 $r'\in I_s^{x'}$,存在 $\hat{x}\in\mathcal{D}$,满足 $\|\hat{x}-x'\|_1=s$ 及 $r'=f(\hat{x})$.再由 $s\le t-2<t-1$,有 $I_t^x\cap I_s^{x'}=\varnothing$. □

记 $I^x=\{I_i^x: i\in\mathbb{N}\}$.引理 1 说明:对任意的邻居 $x,x'\in\mathcal{D}$,每个点 $r\in\mathcal{R}$在 I^x 和 $I^{x'}$中的对应集合序列的位置至多错一位.这个性质为分析机制的差分隐私性提供了极大的方便.并且,这个错位值(即 1)不能再提升到 0 了,否则,或者 $f(x), x\in\mathcal{D}$是恒等函数,或者 $I_0^x=\mathcal{R}, x\in\mathcal{D}$.

另外,$\{I^x: x\in\mathcal{D}\}$是所有满足假设 1 和引理 1 的切分序列中最“瘦”的——从 I_1^x 开始,每个 I_i^x 仅包含了那些与 x 的距离为 i 的邻居的函数值.这个极小性质为分析机制的输出数据有用性提供了方便.

我们的目的是对所有 I_i^x 中的点,赋值以相同的概率密度值.这样,由引理 1 的性质,当判断是否机制满足差分隐私时,只需要判断当$|t-s|\le 1$ 且 $I_t^x\cap I_s^{x'}\ne\varnothing$ 时, $I_t^x, I_s^{x'}$ 中相应点的概率值是否满足差分隐私,而无需考虑其他的点了.

5 线性查询的一种近似最优机制

本节讨论线性查询的最优机制问题.首先讨论当集合序列是极小集合序列$\{I^x: x\in\mathcal{D}\}$时,质量序列$\{\mathcal{Q}^x: x\in\mathcal{D}\}$的最优问题.

线性查询具有特殊的性质:对所有 $x\in\mathcal{D}$及所有 $i\in\mathbb{N}$,集合 $I_i^x-f(x)$ 都是相同的.这是因为:

$$I_0^x-f(x)=\varnothing, I_1^x-f(x)=\{f(x')-f(x): x'\in\mathcal{N}^x\}=\{\pm f(s): s\in\mathcal{X}\},$$

再由 I_i^x 的对称递归构造过程可得.

这个性质为构造对应的值序列提供了方便.将序列$\{I^x: x\in\mathcal{D}\}$代入(替换$\{\mathcal{R}^x: x\in\mathcal{D}\}$)优化问题(3)中,发现对不同的 $x\in\mathcal{D}$,他们有相同的 a_i 和 $b_i, i\in\mathbb{N}$.因此,对所有 $x,y\in\mathcal{D}$,为了让所有目标函数极小值,对质量序列$\{\mathcal{Q}^x: x\in\mathcal{D}\}$,应设 $z_i^x=z_i^y:=z_i, i\in\mathbb{N}$.从而,对所有 $x,y\in\mathcal{D}, \alpha_x=\alpha_y$.这样,优化问题(3)等价于如下的优化问题(5).

$$\left.\begin{array}{l}\underset{z_i,i\in\mathbb{N}}{\arg\min}\left\{\dfrac{\sum_{i=0}^{\infty}\exp(\varepsilon z_i)\int_{r\in I_i^x}\|f(x)-r\|_p\mathrm{d}r}{\sum_{j=0}^{\infty}\exp(\varepsilon z_j)\int_{r\in I_j^x}\mathrm{d}r}:x\in\mathcal{D}\right\}\\ \text{s.t. 对任意邻居}x,x',\text{任意}i,j\in\mathbb{N}\text{且满足}|i-j|\leq 1\text{及}\\ \quad I_i^x\cap I_j^{x'}\neq\varnothing,\text{有}z_i-z_j\leq 1\end{array}\right\}\tag{5}$$

下面分析质量序列$\{z_i:i\in\mathbb{N}\}$的最优取值问题.首先给出两个引理.

引理 **2**. 设 $g(x)=\dfrac{\alpha_0x+\alpha_1}{\beta_0x+\beta_1}$:若$\alpha_0\beta_1>\alpha_1\beta_0$,则 $g(x)$是非减函数;若$\alpha_0\beta_1<\alpha_1\beta_0$,则 $g(x)$是非增函数.

证明:由于 $g(x)$的导数 $g'(x)=\dfrac{\alpha_0\beta_1-\alpha_1\beta_0}{(\beta_0x+\beta_1)^2}$,从而可得结论. □

引理 **3**. 设 $f:\mathcal{X}\to\mathbb{R}^n$ 为一线性查询函数,且设 $a_i=\int_{r\in I_i^x}\|f(x)-r\|_p\mathrm{d}r, b_i=\int_{r\in I_i^x}\mathrm{d}r$,则对 $j>i$,有 $\dfrac{a_i}{b_i}\leq\dfrac{a_j}{b_j}$.

证明:该结论是积分中值定理的一个推论.

由引理 3,可得如下推论。

推论 **1**. 对所有 $t\in\mathbb{N}$,有 $\dfrac{\sum_{i=0}^{t-1}\exp(-\varepsilon i)a_i}{\sum_{i=0}^{t-1}\exp(-\varepsilon i)b_i}\leq\dfrac{a_t}{b_t}$.

定理 **1**. 设假设 1 成立,则优化问题(5)的最优质量序列为 $\{z_i^x=-i\}_{i\in\mathbb{N}},x\in\mathcal{D}$.

证明:本定理使用数学归纳法证明,其证明思路简述如下.

首先,根据推论 1、引理 2 及 z_0=0,可得优化问题(6)的最优解为 z_1=−1.

$$\left.\begin{array}{l}\underset{z_1}{\arg\min}\left\{\dfrac{\sum_{i=0}^{1}\exp(\varepsilon z_i)\int_{r\in I_i^x}\|f(x)-r\|_p\mathrm{d}r}{\sum_{j=0}^{1}\exp(\varepsilon z_j)\int_{r\in I_j^x}\mathrm{d}r}:x\in\mathcal{D}\right\}\\ \text{s.t. 对任意邻居}x,x',\text{任意}i,j\in\{0,1\}\text{满足}\\ \quad I_i^x\cap I_j^{x'}\neq\varnothing,\text{有}z_i-z_j\leq 1\end{array}\right\}\tag{6}$$

接着,设对所有 $t<k$,有 $z_t=-t$.下面证明 $z_k=-k$ 是如下优化问题(7)的解,而这个结论是推论 1 及假设($t<k$ 时,有 $z_t=-t$)的推论.

$$\left.\begin{array}{l}\underset{z_t}{\arg\min}\left\{\dfrac{\sum_{i=0}^{t}\exp(\varepsilon z_i)\int_{r\in I_i^x}\|f(x)-r\|_p\mathrm{d}r}{\sum_{j=0}^{t}\exp(\varepsilon z_j)\int_{r\in I_j^x}\mathrm{d}r}:x\in\mathcal{D}\right\}\\ \text{s.t. 对任意邻居}x,x',\text{任意}i\in\{t-1,t\}\text{满足}\\ \quad I_i^x\cap I_t^{x'}\neq\varnothing,\text{有}z_i-z_t\leq 1\end{array}\right\}\tag{7}$$

□

集合序列$\{I^x:x\in\mathcal{D}\}$是优化问题(3)的极小集合序列,但并不一定是优化问题(3)的最优集合序列.经过本节的研究已经发现:对线性查询问题, $z_t^x=-t,t\in\mathbb{N},x\in\mathcal{D}$ 是(3)的最优质量序列.现在来研究对 $z_t^x=-t,t\in\mathbb{N},x\in\mathcal{D}$ 寻找最优的集合序列,即,满足优化问题(8)的最优解.

$$\begin{aligned}&\underset{\mathcal{R}^x:x\in\mathcal{D}}{\arg\min}\left\{\frac{\sum_{i=0}^{\infty}\exp(-\varepsilon i)\int_{r\in\mathcal{R}_i^x}\|f(x)-r\|_p\mathrm{d}r}{\sum_{j=0}^{\infty}\exp(-\varepsilon j)\int_{r\in\mathcal{R}_j^x}\mathrm{d}r}:x\in\mathcal{D}\right\}\\&\text{s.t. 对任意邻居}x,x',\text{任意}i,j\in\mathbb{N}\text{满足}\\&\quad\mathcal{R}_i^x\cap\mathcal{R}_j^{x'}\neq\varnothing,\text{有}|i-j|\leq 1\end{aligned}\tag{8}$$

我们从极小序列$\{I^x:x\in\mathcal{D}\}$出发来构造最优集合序列$\{\mathcal{R}^x:x\in\mathcal{D}\}$.现在对极小集合序列$\{I^x:x\in\mathcal{D}\}$进行调整以减小目标函数.根据引理2及引理3的结论,可将单位球$\{r\in\mathbb{R}^n:\|f(x)-r\|_p\leq 1\}$中的点添入$I_0^x$(也就是增加了单位球中点的输出概率)以减小目标函数值.但由于$I_t^x=\bigcup_{x'\in\mathcal{N}^x}I_{t-1}^{x'}\setminus A_{t-1}^x,t>0$,对集合$I_0^x$的改变会导致$I_t^x,t>0$相应的变化.这些因素综合的结果对目标函数的增减影响难以用简单的分析判断出来.因此,我们设置一个参数δ,通过解决一个优化问题来决定将哪些点添入I_0^x中.设$\mathcal{R}_0^x=\{r\in\mathbb{R}^n:\|f(x)-r\|_p\leq\delta\},\mathcal{R}_i^x=\bigcup_{x'\in\mathcal{N}^x}\mathcal{R}_{i-1}^{x'}\setminus\bigcup_{j=0}^{i-1}\mathcal{R}_j^x$,我们需要找到最优的$\delta$,使得目标函数达到最小.与集合序列$\{I^x:x\in\mathcal{D}\}$类似,集合序列$\{\mathcal{R}^x:x\in\mathcal{D}\}$也有类似于引理 1 的性质,如引理 4 所示.引理 4 的证明与引理 1 的证明类似,在此不再详述.

引理 **4**. 对任意的邻居$x,x'\in\mathcal{D}$,及对任意的$|t-s|>1$,有$\mathcal{R}_t^x\cap\mathcal{R}_s^{x'}=\varnothing$且$\mathcal{R}=\bigcup_{i=0}^{\infty}\mathcal{R}_i^x$.

根据引理 4 及优化问题(8),可构造无约束优化问题(9),以求解参数δ:

$$\underset{\delta}{\arg\min}\frac{\sum_{i=0}^{\infty}\exp(-\varepsilon i)\int_{r\in\mathcal{R}_i^x}\|f(x)-r\|_p\mathrm{d}r}{\sum_{j=0}^{\infty}\exp(-\varepsilon j)\int_{r\in\mathcal{R}_j^x}\mathrm{d}r}\tag{9}$$

5.1 小 结

下面就概率密度族$\{(\mathcal{Q}^x,\mathcal{R}^x):x\in\mathcal{D}\}$对优化问题(5)的近似最优性进行分析,其中,

$$\mathcal{R}_0^x=\{r\in\mathcal{R}:d(f(x),r)\leq\delta^*\},\mathcal{R}_i^x=\bigcup_{x'\in\mathcal{N}^x}\mathcal{R}_{i-1}^{x'}\setminus\bigcup_{j=0}^{i-1}\mathcal{R}_j^x,z_i^x=-i,i\in\mathcal{N},x\in\mathcal{D}.$$

δ^*为优化问题(9)的最优解.由于δ^*是通过求解优化问题(9)来找到的,因此$\mathcal{R}_i^x$是这种递归构造序列中最优的集合序列.又因为$z_i^x=-i,i\in\mathbb{N},x\in\mathcal{D}$是满足不等式$z_i^x-z_j^{x'}\leq 1,|i-j|\leq 1$且$\mathcal{R}_i^x\cap\mathcal{R}_j^{x'}\neq\varnothing$的最大间距的序列,若将一些点集从$\mathcal{R}_j^x$转入$\mathcal{R}_i^x,j>i$,这样就会增大该点集出现的概率,从而增大目标函数.若将一些点集从$\mathcal{R}_i^x$转入$\mathcal{R}_j^x,j>i$,则就会使得差分隐私不等式约束不满足.并且对任意邻居数据集x,x',密度函数$(\mathcal{Q}^x,\mathcal{R}^x)$与$(\mathcal{Q}^{x'},\mathcal{R}^{x'})$在所有点集$\mathcal{R}_i^x\cap\mathcal{R}_{i-1}^{x'}$都达到了差分隐私不等式$z_i^x-z_{i-1}^{x'}\leq 1$的边界.综上所述,概率密度族$\{(\mathcal{Q}^x,\mathcal{R}^x):x\in\mathcal{D}\}$是优化问题(5)的一个近似 Pareto 最优解.

现在将线性函数相关的最优机制问题归纳为如下的结论.

定理 **2**. 设假设 1 成立.设$f:\mathcal{X}\to\mathbb{R}^n$是任意线性查询函数,则其关于优化问题(2)的一个近似 Pareto 最优解为$\{p^x=(\mathcal{Q}^x,\mathcal{R}^x):x\in\mathcal{D}\}$,其中,

$$\mathcal{Q}^x=\{z_i^x=-i:i\in\mathbb{N}\},\mathcal{R}^x=\{\mathcal{R}_i^x:i\in\mathbb{N}\},\mathcal{R}_0^x=\{r:\|r-f(x)\|_p\leq\delta^*\},\mathcal{R}_i^x=\bigcup_{x'\in\mathcal{N}^x}\mathcal{R}_{i-1}^{x'}\Big/\bigcup_{j=0}^{i-1}\mathcal{R}_j^x$$

及

$$\delta^*=\underset{\delta}{\arg\min}\frac{\sum_{i=0}^{\infty}\exp(-\varepsilon i)\int_{r\in\mathcal{R}_i^x}\|f(x)-r\|_p\mathrm{d}r}{\sum_{j=0}^{\infty}\exp(-\varepsilon j)\int_{r\in\mathcal{R}_j^x}\mathrm{d}r}.$$

注释:本文使用先实现差分隐私再提高有用性的策略来达到隐私和有用性之间的权衡.首先,我们寻找实现差分隐私所需设置的最小点集.具体说来,对数据集x,我们将x的邻居集对应的函数值的集合设置为I_1^x,而

$I_0^x=\{f(x)\}$.依此类推,I_i^x 包含了 I_{i-1}^x 对应的数据集的邻居集合的函数值(需要去掉一些与前面集合的重复点).这样,我们只需要让 I_i^x 集合中的点与 I_{i-1}^x 集合中的点的密度函数值满足差分隐私不等式就可以实现差分隐私了.这种实现差分隐私的方法称为最小点集实现法.在实现差分隐私的基础上,我们接着来提高机制的有用性,这通过引入一个 $f(x)$的δ邻域来实现.具体思路是,让 $f(x)$附近的点获得较高的密度函数值.这就是构造集合序列 $\{\mathcal{R}_i^x:i\in\mathbb{N}\}$ 的原因.这种将差分隐私实现机制通过拆分、组合的方式来实现的方法,可以更精确地对隐私和有用性进行权衡,相关实验也证明了这种方法的合理性.

不过,需要说明的是,本文的方法没能实现真正意义上的机制最优性.最优机制需要找到一个密度函数族 $\{p^x(r):x\in\mathcal{D}\}$,其比其他任何密度函数族都具有更好的数据有用性,而每个密度函数族都具有无穷多条密度函数.这是一个非常复杂的泛函优化问题,且是否存在这种真正意义上的最优机制现在还有疑问.这也是本文以一种近似最优而非最优来限定本文最优性的意图,文献[18]中的不合理结论也充分说明了这个问题.关于差分隐私最优机制的存在性及如何合理定义差分隐私机制最优性,还需要进一步的研究.

6 实例分析

本节通过实际例子来解释本文的方法并分析其性质.

设 $f:\mathcal{X}\to\mathbb{R}$为线性查询函数.对数据集 $x\in\mathcal{D}$,设$\mathcal{N}^x$ 是 x 的所有邻居的集合,则$\mathcal{N}^x$ 对应的函数值的集合为 $\{f'(x):x'\in\mathcal{N}^x\}=f(x)\pm\{f(y):y\in\mathcal{X}\}$.记$\mathcal{V}=\{f(y):y\in\mathcal{X}\}$,则集合$\mathcal{V}$由该查询函数唯一确定.我们称$\mathcal{V}$为查询函数的邻居集,我们下面研究集合$\mathcal{V}$的结构对查询结果的影响.显然,$\mathcal{V}$对应查询函数的局部敏感度和全局敏感度都相同,且等于$\mathcal{V}$中最大元素和最小元素相差的值,即 $\Delta f=\max_{a,b\in\mathcal{V}}|a-b|$.本文的算法试图通过研究$\mathcal{V}$的内部结构来达到构造低噪声差分隐私机制的目的.

考虑$\mathcal{V}$是两个不同区间并的情形,即$\mathcal{V}=[a,b]\cup[b,c]$,其中,$0<a<b<c$.显然,敏感度$\Delta f=c$ 且$\mathcal{V}$的 Lebesgue 测度$\mu(\mathcal{V})=a+c-b$.我们试图通过对 a,b,c 赋以不同的值,从而控制Δf 及$\mu(\mathcal{V})$的值,然后研究其对噪声复杂度的影响.对本节的特例,本文构造差分隐私机制的算法见算法 1 所示,其输出 $f(x)=0$ 对应数据集 x 的密度函数.对应密度函数给所有$\mathcal{R}_i$中的点赋以相同的密度值 $p_i(r)=\exp(-i\varepsilon)/\alpha$.

算法 **1**. 生成概率密度函数.

输入:3 个递增的正实数 a,b,c,正实数δ及ε,迭代次数 n;

//设集合序列$\{\mathcal{R}_i:i\in\mathbb{N}\}$在第 n 步后收敛;

输出:概率密度函数$\{(p_i(r),\mathcal{R}_i):i\in\mathbb{N}\}$.

步骤 1:设置 $I_0=\{0\}$;

步骤 2://计算集合序列 I_i

For i=1 to n {

对每个区间$[s,e]\in I_{i-1}$ {

将 4 个区间$[s,e+a],[s+b,e+c],[s-a,e],[s-c,e-b]$存入集合 I_i 中;

}

计算 $I_i=I_i-\bigcup_{j=0}^{i-1}I_j$; //去掉 I_i 中与前面集合重复点集;

}

步骤 3: //计算集合序列$\mathcal{R}_i$

For i=0 to n {

对每个区间$[s,e]\in I_i$ {

将区间$[s-\delta,e+\delta]$存入集合$\mathcal{R}_i$中;

}

计算 $\mathcal{R}_i=\mathcal{R}_i-\bigcup_{j=0}^{i-1}\mathcal{R}_j$; //去掉$\mathcal{R}_i$中与前面集合重复点集;

}

//设$\mathcal{R}_n$={[$-a_n$,$-\Delta f$,$-a_n$],[a_n,$a_n+\Delta f$]}

步骤 4:计算 $sum1=\sum_{i=0}^{n}\exp(-i\varepsilon)\mu(\mathcal{R}_i)$ 及 $sum2=2\Delta f*\exp(-\varepsilon(n+1))/(1-\exp(-\varepsilon))$;

步骤 5:计算$\alpha=sum1+sum2$;

步骤 6:设置 $p_i(r)=\exp(-i\varepsilon)/\alpha$,其中,$r\in\mathcal{R}_i$,$i\in\mathbb{N}$;

步骤 7:Return $\{(p_i(r),\mathcal{R}_i):i\in\mathbb{N}\}$.

算法 1 需要下面的集合序列收敛性质:

定义 **2**. 设$\mathcal{R}=\mathbb{R}$.对数据集 $x\in\mathcal{D}$,称集合序列$\{\mathcal{R}_i:i\in\mathbb{N}\}$在第 $n\in\mathbb{N}$步收敛,如果存在 $a_n\in\mathcal{R}$,有:

$$\mathcal{R}_n=\pm(a_n+[0,\Delta f])\text{且}\mathcal{R}_{n+1}=\pm(a_n+[\Delta f,2\Delta f]).$$

我们的实验设计如下:首先,计算算法 1 生成的密度函数的数学期望值 $mean=\sum_{i=0}^{\infty}p_i\times\int_{r\in\mathcal{R}_i}r\mu(\mathrm{d}r)$; 然后,与 Laplace 机制[27]的期望值$\Delta f/\varepsilon$及 Staircase 机制[18]的期望值$\Delta f\times\exp(\varepsilon/2)/(\exp(\varepsilon)-1)$进行比较,越小的期望值代表噪声复杂度越低.由于发现 Staircase 机制的期望值比 Laplace 机制的期望值小,因此,本实验只比较 *mean* 与 Staircase 机制期望值$\Delta f\times\exp(\varepsilon/2)/(\exp(\varepsilon)-1)$的大小,即 $rate=mean\times(\exp(\varepsilon)-1)/(\Delta f\times\exp(\varepsilon/2))$.*rate* 值小于 1,代表算法 1 对应机制优于 Staircase 机制;反之,则差于 Staircase 机制.

我们构造两组实验:第 1 组使用相同的敏感度及不同的函数邻居集测度,第 2 组使用不同的敏感度及相同的函数邻居集测度.第 1 组实验有 4 个不同的查询函数 f_1,f_2,f_3,f_4,其邻居集分别为$\mathcal{V}_1$=[0,1]∪[1000,1001],$\mathcal{V}_2$=[0,100]∪[1000,1001],$\mathcal{V}_3$=[0,500]∪[1000,1001],$\mathcal{V}_4$=[0,1001].4 个不同集合可以理解为 4 个不同单位的工资分布情况:$\mathcal{V}_1$代表了工资两级分化极其严重的单位;$\mathcal{V}_4$代表了该单位的工资虽然最高工资和最低工资差距很大,但是高中低档工资都可以出现;其他两个集合代表了折中的情形.实验结果如图 1 所示.

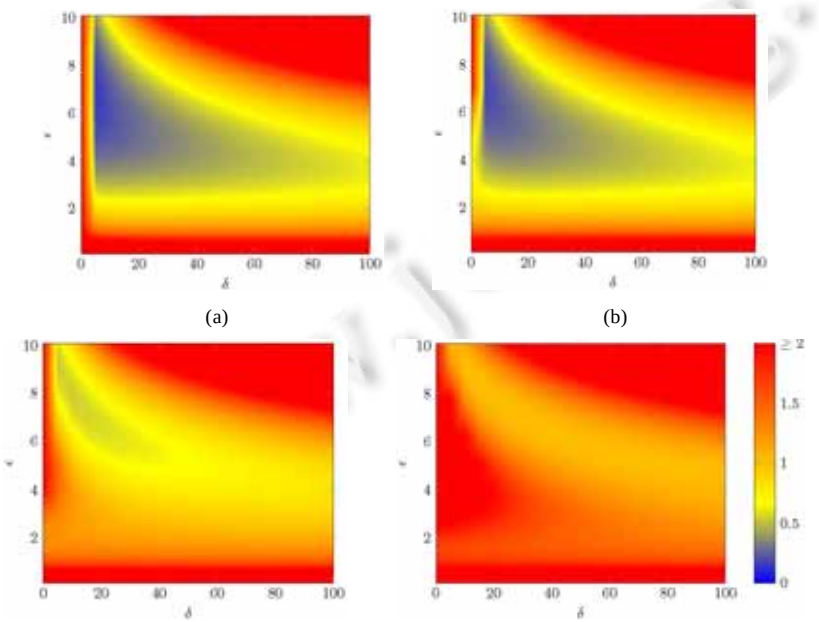
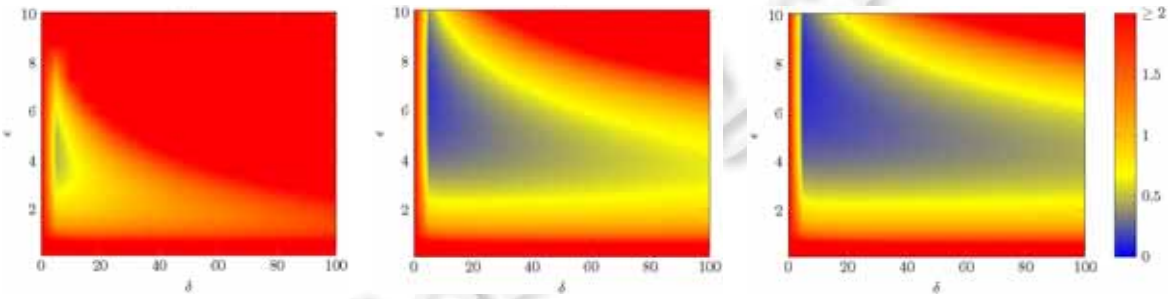


Fig.1

图 1

图 1 的 4 个子图分别代表了$\mathcal{V}_1,\mathcal{V}_2,\mathcal{V}_3,\mathcal{V}_4$的实现效果图.其中,横坐标表示$\delta$的取值,纵坐标表示$\varepsilon$的取值,坐标点$(\delta,\varepsilon)$对应的值表示相应的 *rate* 的值.从图 1 可以发现:随着邻居集测度$\mu(\mathcal{V}_i)$的增大,*rate* 的值相应增大,从而说明算法 1 中机制的优势是随$\mu(\mathcal{V}_i)$的增大而递减的.

第 2 组实验有 3 个查询函数,其邻居集分别为$\mathcal{V}_5$=[0,1]∪[100,101],$\mathcal{V}_6$=[0,1]∪[1000,1001],$\mathcal{V}_7$=[0,1]∪[2000,2001].显然,3 个函数的全局(及局部)敏感度分别为 101,1001,2001,但有相同的邻居集测度$\mu(\mathcal{V}_i)$=2.其实验结果如图 2 所示.



(a) (b) (c)

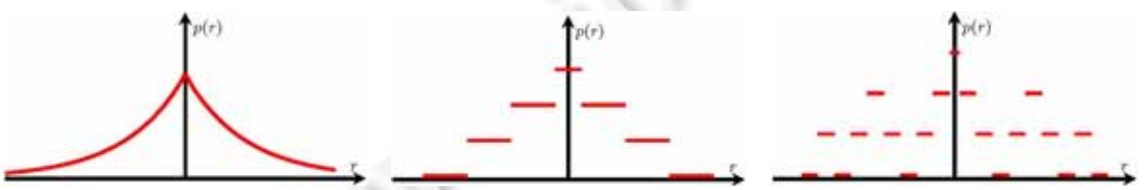
Fig.2

图 2

图 2 的 3 个子图分别代表了$\mathcal{V}_5,\mathcal{V}_6,\mathcal{V}_7$的实现效果图.与图 1 一样,每个子图中坐标点$(\delta,\varepsilon)$对应的值表示相应的 *rate* 的值.从图 2 可以发现:当敏感度增加的时候,算法 1 中机制的噪声复杂度要比 Staircase 机制的复杂度相对越来越小.也就是说:在函数邻居集的测度不变化的条件下,随着敏感度的增加,算法 1 中机制的噪声复杂度要比 Staircase 等敏感度方法的噪声复杂度越来越小.

综合图 1、图 2 的结果,我们可以得到如下的结论:若查询函数的邻居集$\mathcal{V}$的敏感度Δf与其测度$\mu(\mathcal{V})$的比率$\Delta f/\mu(\mathcal{V})$越来越大时,敏感度方法将添加越来越大的噪声.而本文的方法会极大地降低这个比率带来的噪声复杂度升高的问题,这也是本文的方法优于敏感度方法的最显著特征.

Laplace 机制、Staircase 机制及算法 1 中机制的密度函数如图 3 所示.图 3(c)的密度函数与前两个密度函数的显著区别是:密度函数并不是从中心向两边递减的,而是总体递减,但是局部有增的密度函数.这种密度函数更适合于$\Delta f/\mu(\mathcal{V})$很大的情形,因为其波浪状形态可以不受全局敏感度的限制而更适合于邻居集的(不同查询函数的)多变特性.



(a) Laplace 机制 (b) Staircase 机制 (c) 算法 1 中机制

Fig.3 The density functions of Laplace mechanism, Staircase mechanism and the mechanism in algorithm 1

图 3 Laplace 机制、Staircase 机制、算法 1 中机制的密度函数

算法 2 是算法 1 生成的密度函数的随机变量生成算法.

算法 **2**. 生成随机变量.

输入:敏感度Δf、迭代次数 n、$\{\mathcal{R}_i:i\in\{0,1,\ldots,n\}\}$、$a_n$、$\alpha=sum1+sum2$;

//设$\mathcal{R}_n=\{[-a_n,-\Delta f,-a_n],[a_n,a_n+\Delta f]\}$,即,算法从第 n 步后收敛;

输出:随机变量 X.

步骤 1:以 $\exp(-i\varepsilon)\mu(\mathcal{R}_i)/\alpha$的概率输出 $i,i\in\{0,1,\ldots,n\}$;以概率 $sum2/\alpha$输出 $n+1$;

步骤 2:If 步骤 1 输出值 $i\in\{0,1,\ldots,n\}$ then

在$\mathcal{R}_i$ 中均匀抽样出一个数 x;

Return x;

Else

从参数为 $1-\exp(-\varepsilon)$的几何分布中抽样出一个整数 j;

从$[-a_n-(j+1)\Delta f,-a_n-j\Delta f]\cup[a_n+j\Delta f,a_n+(j+1)\Delta f]$中均匀随机抽样出一个数 y;

Return y;

6.1 时间复杂度

算法 1 和算法 2 的时间复杂度由集合序列$\{\mathcal{R}_i:i\in\mathbb{N}\}$的收敛性决定.本节的例子中,该序列都在 2 200 步内达到了收敛.但是对一般线性查询函数,我们还没能有方法证明其一定在有限步内收敛.不过,我们相信这个结论是对的.

若$\{\mathcal{R}_i:i\in\mathbb{N}\}$在第 n 步收敛,且 $\max_{i\in\{0,1,\ldots,n\}}N_i=N$,其中,$N_i$ 是$\mathcal{R}_i$ 中区间的个数,则算法 1 的时间复杂度为 $O(n^2\times N^2)$,算法 2 的时间复杂度为 $O(n)$.

7 结 论

传统的观点认为,查询函数的(全局或局部)敏感度是查询函数噪声复杂度的(最主要)标志.本文的结论发现:敏感度其实只是查询函数的邻居集$\mathcal{V}^x$ 的一个极值特征,还有很多敏感度无法刻画的特征(如该集合中点的分布).本文的方法可根据$\mathcal{V}^x$ 中点的分布情形,构造适合于该分布的差分隐私机制,相应的密度函数类似于图 3(c)所示.与敏感度方法(如图 3(a)、图 3(b)所示)不同,本文机制的密度函数不是(向两边)全局递减的,而是以局部有起伏的方式(向两边)递减的,这种递减方式更具灵活性,更适合于$\mathcal{V}^x$ 中点的不规则分布情形.

第 6 节的实际例子也说明,本文的方法一般要比 Laplace 机制及 Staircase 机制更精确.但是本文的方法具有很高的时间复杂度,不适合于直接使用在大数据集上.如何通过本文的方法和结论构造低时间复杂度的非敏感度方法,将作为未来的一项工作.

另外,本文的很多分析方法和结论(如第 3 节、第 4 节中的内容都没有设定函数类型)也可对非线性查询问题[15,16,34−36]进行分析,因为非线性查询问题具有更加复杂的邻居集.至于其分析有效程度如何,还需要进一步探索.非线性查询问题将具有更加复杂的最优机制,不同数据集将对应形状各异的密度函数,对非线性函数的非敏感度机制研究将作为未来的另一项工作.

致谢 本文作者感谢匿名审稿专家对本文初稿提出的宝贵建议和意见,这些建议和意见对本文的完整性及易读性有很大的帮助.同时,这些建议和意见促使我们发现了初稿中的一个证明错误.

References:

[1] Ganta SR, Kasiviswanathan SP, Smith A. Composition attacks and auxiliary information in data privacy. In: Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2008. 265−273. [doi: 10.1145/1401890.1401926]

[2] Narayanan A, Shmatikov V. Robust de-anonymization of large sparse datasets. In: Proc. of the 2008 IEEE Symp. on Security and Privacy (S&P 2008). 2008. 111−125. [doi: 10.1109/SP.2008.33]

[3] Aggarwal CC, Yu PS. Privacy-Preserving data mining—Models and algorithms. In: Proc. of the Advances in Database Systems, Vol.34. Springer-Verlag, 2008. [doi: 10.1007/978-0-387-70992-5]

[4] Dwork C. A firm foundation for private data analysis. Communications of the ACM, 2011,54(1):86−95. [doi: 10.1145/1866739.1866758]

[5] Dwork C, Roth A. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 2014,9(3-4):211−407. [doi: 10.1561/0400000042]

[6] Jain P, Thakurta A. Differentially private learning with kernels. In: Proc. of the 30th Int'l Conf. on Machine Learning. 2013. 118−126.

[7] Zhang J, Zhang ZJ, Xiao XK, Yang Y, Winslett M. Functional mechanism: Regression analysis under differential privacy. Proc. of the VLDB Endowment, 2012,5(11):1364−1375. [doi: 10.14778/2350229.2350253]

[8] Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao XK. Privbayes: Private data release via bayesian networks. In: Proc. of the 2014 ACM SIGMOD Int'l Conf. on Management of Data. 2014. 1423−1434. [doi: 10.1145/2588555.2588573]

[9] Li NH, Qardaji W, Su D. Provably private data anonymization: Or, *k*-anonymity meets differential privacy. CERIAS Tech Report, 2010.

[10] Soria-Comas J, Domingo-Ferrer J, Sánchez D, Martínez S. Enhancing data utility in differential privacy via microaggregation-based *k*-anonymity. The Int'l Journal on Very Large Data Bases, 2014,23(5):771−794. [doi: 10.1007/s00778-014-0351-4]

[11] Kasiviswanathan SP, Lee HK, Nissim K, Raskhodnikova S, Smith AD. What can we learn privately? In: Proc. of the IEEE 49th Annual IEEE Symp. on Foundations of Computer Science. 2008,40(3):793−826. [doi: 10.1109/FOCS.2008.27]

[12] Wasserman L, Zhou S. A statistical framework for differential privacy. Journal of the American Statistical Association, 2010, 105(489):375−389. [doi: 10.1198/jasa.2009.tm08651]

[13] Chaudhuri K, Hsu D. Convergence rates for differentially private statistical estimation. In: Proc. of the Int'l Conf. on Machine Learning. 2012. 1327−1334.

[14] Kellaris G, Papadopoulos S, Xiao XK, Papadias D. Differentially private event sequences over infinite streams. Proc. of the VLDB Endowment, 2014,7(12):1155−1166. [doi: 10.14778/2732977.2732989]

[15] Karwa V, Raskhodnikova S, Smith AD, Yaroslavtsev G. Private analysis of graph structure. ACM Trans. on Database Systems, 2014,39(3):22:1−22:33. [doi: 10.1145/2611523]

[16] Chaudhuri D, Sarwate AD, Sinha K. A near-optimal algorithm for differentially-private principal components. Journal of Machine Learning Research, 2013,14(1):2905−2943.

[17] Nikolov A, Talwar D, Zhang L. The geometry of differential privacy: The sparse and approximate cases. In: Proc. of the Annual ACM Symp. on Theory of Computing. 2013. 351−360. [doi: 10.1145/2488608.2488652]

[18] Geng Q, Viswanath P. The optimal noise-adding mechanism in differential privacy. IEEE Trans. on Information Theory, 2016, 62(2):925−951. [doi: 10.1109/TIT.2015.2504967]

[19] Gupte M, Sundararajan M. Universally optimal privacy mechanisms for minimax agents. In: Proc. of the 29th ACM SIGMOD-SIGACT-SIGART Symp. on Principles of Database Systems. 2010. 135−146. [doi: 10.1145/1807085.1807105]

[20] Li C, Miklau G. Optimal error of query sets under the differentially-private matrix mechanism. In: Proc. of the 16th Int'l Conf. on Database Theory. 2013. 272−283. [doi: 10.1145/2448496.2448529]

[21] Yaroslavtsev G, Cormode G, Procopiuc CM, Srivastava D. Accurate and efficient private release of datacubes and contingency tables. In: Proc. of the IEEE 29th Int'l Conf. on Data Engineering (ICDE). 2013. 745−756. [doi: 10.1109/ICDE.2013.6544871]

[22] Wang ZT, Fan K, Zhang JQ, Wang LW. Efficient algorithm for privately releasing smooth queries. In: Proc. of the Advances in Neural Information Processing Systems. 2013. 782−790.

[23] Geng Q, Viswanath P. Optimal noise adding mechanisms for approximate differential privacy. IEEE Trans. on Information Theory, 2016,62(2):952−969. [doi: 10.1109/TIT.2015.2504972]

[24] Brenner H, Nissim K. Impossibility of differentially private universally optimal mechanisms. In: Proc. of the 51st IEEE Annual Symp. on Foundations of Computer Science. 2010. 71−80[doi: 10.1109/FOCS.2010.13]

[25] Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao XK. Private release of graph statistics using ladder functions. In: Proc. of the 2015 ACM SIGMOD Int'l Conf. on Management of Data. 2015. 731−745. [doi: 10.1145/2723372.2737785]

[26] Raskhodnikova S, Smith AD. Smooth sensitivity and sampling in private data analysis. In: Proc. of the 39th Annual ACM Symp. on Theory of Computing. 2007. 75−84. [doi: 10.1145/1250790.1250803]

[27] Dwork C. Differential privacy. In: Proc. of the Int'l Colloquium on Automata, Languages, and Programming, 2006. 1−12. [doi: 10.1007/978-3-540-79228-4_1]

[28] Hardt M, Talwar K. On the geometry of differential privacy. In: Proc. of the 42nd ACM Symp. on Theory of Computing. 2010. 705−714. [doi: 10.1145/1806689.1806786]

[29] Ghosh A, Roughgarden T, Sundararajan M. Universally utility-maximizing privacy mechanisms. In: Proc. of the 41st ACM Symp. on Theory of Computing. 2009. 351−360. [doi: 10.1145/1536414.1536464]

[30] Gupta A, Roth A, Ullman J. Iterative constructions and private data release. In: Proc. of the 9th Int'l Conf. on Theory of Cryptography. 2012. 339−356. [doi: 10.1007/978-3-642-28914-9_19]

[31] Roth A, Roughgarden T. Interactive privacy via the median mechanism. In: Proc. of the 42nd ACM Symp. on Theory of Computing. 2010. 765−774. [doi: 10.1145/1806689.1806794]

[32] Hardt M, Ligett K, McSherry F. A simple and practical algorithm for differentially private data release. In: Proc. of the Advances in Neural Information Processing Systems. 2010. 2339−2347.

[33] Collette Y, Siarry P. Multiobjective Optimization: Principles and Case Studies. Springer Science & Business Media, 2013. 15−43.

[34] Lu GQ, Zhang XJ, Ding LP, Li YF, Liao X. Frequent sequential pattern mining under differential privacy. Journal of Computer Research and Development, 2015,52(12):2789−2801 (in Chinese with English abstract) [doi: 10.7544/issn1000-1239.2015.20140516]

[35] Ouyang J, Yin J, Liu SP. Differential privacy publishing strategy for distributed transaction data. Ruan Jian Xue Bao/Journal of Software, 2015,26(6):1457−1472 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4526.htm [doi: 10.13328/j.cnki.jos.004576]

[36] Chen SX. New methods for linear queries in providing differential privacy protection [Ph.D. Thesis]. Shanghai: Fudan University, 2012 (in Chinese with English abstract).

附中文参考文献:

[34] 卢国庆,张啸剑,丁丽萍,李彦峰,廖鑫.差分隐私下的一种频繁序列模式挖掘方法.计算机研究与发展,2015,52(12):2789−2801. [doi: 10.7544/issn1000-1239.2015.20140516]

[35] 欧阳佳,印鉴,刘少鹏.一种分布式事务数据的差分隐私发布策略.软件学报,2015,26(6):1457−1472. http://www.jos.org.cn/1000-9825/4576.htm [doi: 10.13328/j.cnki.jos.004576]

[36] 陈世熹.提供差分隐私保护的线性查询新方法[博士学位论文].上海:复旦大学,2012.

武跟强(1980－),男,甘肃天水人,博士生,CCF 学生会员,主要研究领域为安全多方计算与隐私保护,差分隐私理论.

贺也平(1962－),男,博士,研究员,博士生导师,主要研究领域为系统安全,隐私保护.

夏娴瑶(1986－),女,博士生,CCF 学生会员,主要研究领域为隐私保护.