差分隐私下多重一致性约束问题的逼近方法

蔡剑平[1]，刘西蒙[1]，熊金波[2]，应作斌[3]，吴英杰[1]

（1. 福州大学数学与计算机科学学院，福建 福州 350108；2. 福建师范大学数学与信息学院，福建 福州 350117；
3. 新加坡南洋理工大学电气与电子工程学院，新加坡 639798）

摘 要：为了解决差分隐私下多重一致性约束的最优发布问题，通过分析最优一致性发布原理提出了多重一致性约束问题的逼近方法。所提方法的主要思想是将一致性约束问题划分为多个一致性约束子问题，通过反复独立地求解各一致性约束子问题实现原问题的最优一致性发布。其优势在于一致性约束问题划分之后，子问题往往更容易求解或者实现子问题最优一致性发布的技术已相当成熟，从而能够解决更加复杂的差分隐私最优发布问题。分析论证了逼近方法的收敛性，保证任意一致性约束子问题的划分均能实现原问题的最优一致性发布。并且，以销量直方图发布为例，基于多重一致性约束问题的逼近方法设计了差分隐私餐馆销量直方图一致性并行发布算法。实验表明，该算法相比通用解法可提升效率高达 400 倍，并且具备处理百万级大规模数据的能力。

关键词：差分隐私；一致性约束；逼近方法；收敛性；并行计算

中图分类号：TP309

文献标识码：A

DOI: 10.11959/j.issn.1000−436x.2021122

Approximation method of multiple consistency constraint under differential privacy

CAI Jianping[1], LIU Ximeng[1], XIONG Jinbo[2], YING Zuobin[3], WU Yingjie[1]

1. College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China
2. College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350117, China
3. School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore

Abstract: Under differential privacy, to solve the optimal publishing problem with multiple consistency constraints, an approximation method of multiple consistency constraints was proposed by the theoretical analysis of the principle of optimal consistency release. The main idea was to divide the consistency constraint problem into several consistency constraint sub-problems and then achieve the original problem's optimal consistency release by solving each consistency constraint sub-problem repeatedly and independently. The advantage was that after the consistency constraint problem divided, the sub-problems were often easier to solve, or the technology to achieve optimal and consistent release of sub-problems is quite mature. Therefore more complex differential privacy optimal release problem could be solved. After analysis, the approximation method's convergence was fully demonstrated, ensuring that any partition of consistency constrained sub-problems can always achieve the optimal consistency release of the original problem. Furthermore, taking the sales histogram publishing as an example, based on the approximation method of multiple consistency constraints, a parallel algorithm was designed with optimal consistency release under differential privacy. The experimental results show that the algorithm's efficiency is 400 times higher than that of the general solution, and the algorithm can process millions of large-scale data.

Keywords: differential privacy, consistency constraint, approximation method, convergence, parallel computing

---

收稿日期：2021−01−04；修回日期：2021−04−03

通信作者：刘西蒙，snbnix@gmail.com

基金项目：国家自然科学基金资助项目（No.62072109, No.U1804263, No.61702105, No.62002062）

Foundation Item: The National Natural Science Foundation of China (No. 62072109, No.U1804263, No.61702105, No.62002062)

1 引言

随着互联网和大数据技术的快速发展，人们普遍意识到数据发布、信息共享的重要性。为了达到信息公开、成果展示或者提升商业价值、履行社会责任等目的，包括医疗、餐饮、教育、金融等在内的多个行业都在尝试利用互联网和大数据技术向社会公开发布统计信息。然而，数据挖掘技术的飞速发展也使人们从公开发布的信息中挖掘潜在信息的能力不断增强。除了有价值的合法信息外，可被挖掘的信息中还潜藏着大量的个人敏感信息。利用公开发布的信息，攻击者可结合相关背景知识，利用关联分析等技术手段推断或窃取个人隐私，给人们的隐私安全带来了巨大威胁。为保护个人隐私，Dwork[1]提出了差分隐私保护技术。该技术通过对数据添加扰动的方式实现隐私保护，从理论上保证了具备任何知识的攻击者都无法从被保护的公开数据中挖掘个人隐私，是目前公认有效的隐私保护技术。

在某些数据发布问题中，数据间满足了某种语义上的一致性约束。由于差分隐私通过向数据添加噪声实现隐私保护，噪声的随机性会彻底破坏数据间一致性约束。为了获得满足一致性约束的发布效果，不少文献针对各类模型提出了有效的一致性发布算法[2-8]。然而，多数算法所适用的场景针对性较强，难以有效地应用于更广泛的差分隐私最优一致性发布问题。

随着差分隐私技术的日益普及，数据发布场景越来越复杂，同时数据间一致性约束问题的解决难度也越来越高。不少问题已超出了现有技术的关注范围，虽然采用基于极大似然估计的通用解法能够实现最优一致性发布，但通用解法效率极低，无法满足较大规模的数据发布需求。以餐饮行业的直方图统计发布为例，假设某餐馆记录了开业以来所有顾客的消费情况。记录内容如表 1 所示，包含了顾客标识、食物种类、消费数量以及消费时间。假设餐馆提供的食物单品为可乐、汉堡、鸡翅、薯条、鸡块 5 种，则顾客所能购买的食物单品种类数 $c=5$。为了向公众展示销售情况，餐馆决定按天统计销量并采用直方图发布技术公开发布各单品销量，发布内容如表 2 所示。同时，为保护顾客隐私，餐馆决定采用差分隐私技术并希望获得最优的一致性发布效果。本文称发布过程中涉及的问题为餐馆销量直方图发布问题。考虑餐馆长期以来只以套餐的形式销售食物，并不以单品形式销售，导致某些单品的销量组合无法满足该场景的语义特征。例如，餐馆销售以下 3 种套餐：1) 2 份可乐+汉堡+鸡翅+鸡块，2) 2 份汉堡+2 份鸡翅+薯条，3) 可乐+汉堡+薯条+鸡块。

表 1 消费记录

顾客	食物	消费数量/份	时间
Alice	可乐	1	Day1
Bob	汉堡	1	Day1
⋮			
Tom	可乐	2	Day25
⋮			

表 2 销量统计

食物	销量/份				
	Day1	Day2	…	Day25	…
可乐	35	25	…	31	…
汉堡	44	28	…	37	…
鸡翅	31	21	…	26	…
薯条	23	13	…	19	…
鸡块	24	16	…	21	…

在上述套餐组合下某些销量数据不可能出现。例如，不会出现 5 种单品的销量均为 3 份的情况。因此，除了直方图一致性约束问题，餐馆销量直方图发布问题还面临由套餐组合导致的一致性约束问题。本文称该问题为套餐一致性约束问题。餐馆销量直方图发布问题是由两者共同组成的全局一致性约束问题，超出了直方图发布问题的研究范畴，更加复杂。而一致性约束子问题的求解容易得多。直方图一致性约束问题的研究[3-5]已相当成熟，现有技术已具备实现大规模直方图最优一致性发布的能力；并且，由于餐馆提供的食物单品仅有 5 种，每个套餐一致性约束子问题仅为数据规模为 5 的小问题，采用通用解法[6]即可高效求解。数据规模小或存在高效算法等原因常常使子问题的求解难度比全局问题容易得多，但子问题之间并非简单的叠加关系，分别解决子问题的结果往往不能解决全局问题。研究表明，单独对某个子问题执行最优一致性发布算法会破坏另一个子问题发布的一致性。

为充分发挥一致性约束子问题高效求解的优势，提升全局最优一致性问题的求解效率，本文基于最优一致性发布问题的理论分析提出了差分隐私多重一致性约束最优发布问题。该问题主张将复杂的差分隐私的最优一致性约束问题拆分成多个

可高效求解的子问题，然后通过独立求解子问题的最优一致性发布高效地实现原问题的最优一致性发布。经研究，本文提出了差分隐私下多重一致性约束问题的逼近方法，简称多重一致性约束逼近方法。该方法通过反复迭代求解一致性约束子问题使发布结果逼近原问题的最优一致性发布。严格理论论证表明，无论子问题被如何划分，该方法总能保证多重一致性约束问题实现最优一致性发布。此外，不少同类子问题所涉及的发布数据互不相交，数据的发布过程具有独立性，利用这种独立性设计的并行算法可进一步提升多重一致性约束逼近方法的求解效率。

本文主要的研究工作如下。

1) 求得差分隐私下最优一致性发布的解析表达式，并深入分析了解析表达式的数学性质。

2) 基于解析表达式的分析提出差分隐私多重一致性约束问题的逼近方法，并对该方法的收敛性进行充分论证。

3) 讨论多重一致性约束逼近方法的可并行性。以餐馆销量直方图发布问题为例设计了最优一致性发布算法并进行实验分析。

2　相关工作

自 Dwork[1]提出差分隐私以来，不少国内外学者对数据发布的一致性约束问题做了深入研究，提出了许多有效的最优一致性发布算法。其中，以树为基础的发布模型是差分隐私一致性约束问题的典型代表。为解决直方图发布过程中的数据不一致问题，Boosting 算法[2]通过对完全 k 叉树的后置处理实现了最优一致性发布。基于 Boosting 算法，Cormode 等[8]针对空间数据的划分发布问题建立了四分树发布模型，并提出了满足一致性约束的 Quad-Post 算法。考虑 Boosting 算法只能针对完全 k 叉树的不足，吴英杰等[3]提出了 LBLUE（local best linear unbiased estimation）算法实现了面向任意区间树的最优一致性发布，贾俊杰等[5]则通过将查询区间映射为完全 k 叉树的方法改进最优一致性发布。与其他算法不同，LBLUE 算法将区间树中每对父子节点间的等式关系作为一个一致性约束子问题，然后采用迭代逼近的思想求解最优一致性发布。实际上，LBLUE 算法所解决的问题是多重一致性约束最优发布问题的一个特例，其有效性可以通过本文提出的理论得以充分解释。因此，该算法可以视为多重一致性约束逼近方法的一个具体应用。相比于 Boosting 算法，LBLUE 算法不再局限于完全树模型，表明多重一致性约束逼近方法具备了处理更复杂模型的能力。

通过构造虚拟节点，张双越等[7]发现了差分隐私轨迹流量发布过程中潜在的一致性约束问题，通过实现最优一致性发布有效地提升了数据发布的精确性。该结果表明，除了以树为基础的发布模型，差分隐私一致性约束问题还具有其他更多的表现形式。多种不同的差分隐私一致性约束子问题可能存在于一个复杂的发布场景中。然而，目前关于差分隐私一致性约束问题的研究主要针对某个特定的应用场景。虽然大多数一致发布算法都是高效的，但仍无法解决复杂发布场景所涉及的一致性约束问题。采用极大似然估计的思想，Lee 等[6]将差分隐私一致性约束问题表述为抽象的优化方程，并实现了适用于任意最优一致性约束问题的通用解法。然而，通用解法实现最优一致性发布的效率普遍较低，只能有效解决局部的或规模较小的一致性约束问题。如何合理利用高效但针对性强的一致性发布算法以及低效但通用的一致性发布算法解决更复杂的差分隐私最优一致性约束问题具有较高的研究价值。此外，目前多数差分隐私一致性约束问题的研究工作集中在发布精度或效率的提升上。关于最优一致性发布性质的研究还十分有限，现有理论难以解释多重一致性约束之间的内在联系。因此，差分隐私多重一致性约束问题仍存在较大的研究空间。

3　预备知识

3.1　差分隐私

为避免隐私泄露，差分隐私技术通过向待发布数据添加噪声的方式实现隐私保护。通过添加噪声，差分隐私有效地隐藏了隐私信息的存在性，确保攻击者即使掌握了所有背景知识也无法有效推断个人隐私。差分隐私的形式化定义如下。

定义 1　差分隐私[9]。若一个随机算法 $\mathcal{M}$ 满足 $(\varepsilon,\delta)-$差分隐私，则对于 2 个兄弟数据集 D 和 D' 满足所有 $\mathcal{M}$ 的输出 $O\subseteq\text{Range}(\mathcal{M})$ 都有以下不等式成立。

$$\Pr\left(\mathcal{M}(D)\in O\right)\leqslant e^{\varepsilon}\Pr\left(\mathcal{M}(D')\in O\right)+\delta \quad (1)$$

假设待发布数据由 n 个数据组成，分别记为 $x_1,x_2,\cdots,x_n$，则数值型的发布函数为 $\mathcal{A}:\mathcal{D}\to\mathbb{R}^n$。不妨将这些数据依次写为列向量 $\boldsymbol{x}=[x_1,x_2,\cdots,x_n]^{\mathrm{T}}$

的形式，满足 $\boldsymbol{x}=\mathcal{A}(D)$。随机算法 $\mathcal{M}$ 通常采用特定的噪声机制向 $\mathcal{A}(D)$ 添加噪声以实现差分隐私。常见的噪声机制主要包括拉普拉斯机制和高斯机制，其定义分别如下。

定义 2 拉普拉斯机制[10]。对于发布函数 $\mathcal{A}$: $\mathcal{D}\to\mathbb{R}^n$，拉普拉斯机制通过式(2)实现 $(\varepsilon,0)-$差分隐私。

$$\mathcal{M}(D)=\mathcal{A}(D)+\boldsymbol{\xi}=\boldsymbol{x}+\boldsymbol{\xi} \tag{2}$$

其中，$\boldsymbol{\xi}$ 为随机向量且各元素均符合拉普拉斯分布，即 $\xi_i\sim\mathrm{Lap}\left(\frac{\Delta_1}{\varepsilon}\right)$，$\Delta_1$ 为 $\mathcal{A}$ 的 L_1-敏感度[10]。

定义 3 高斯机制[9]。对于发布函数 $\mathcal{A}:\mathcal{D}\to\mathbb{R}^n$，高斯机制通过式(3)实现 $(\varepsilon,\delta)-$差分隐私。

$$\mathcal{M}(D)=\mathcal{A}(D)+\boldsymbol{\xi}=\boldsymbol{x}+\boldsymbol{\xi} \tag{3}$$

其中，$\boldsymbol{\xi}$ 为随机向量且各元素均符合高斯分布，即 $\xi_i\sim\mathcal{N}\left(0,\left(\frac{1+\sqrt{2\ln(1/\delta)}}{\varepsilon}\Delta_2\right)^2\right)$，$\Delta_2$ 为 $\mathcal{A}$ 的 L_2-敏感度。

根据上述定义可知，无论采用拉普拉斯机制还是高斯机制，$\mathcal{M}(D)$ 中添加的噪声都具有独立同分布的性质。由于噪声随机性，$\mathcal{M}(D)$ 无法仅靠噪声机制保证满足任何一致性约束。

3.2 数据发布的一致性约束问题

在数据发布过程中，一致性约束问题是由 m 个一致性约束条件组成的发布问题，其发布结果要求这 m 个一致性约束条件同时满足。其中，一致性约束条件的定义如下。

定义 4 一致性约束条件。对于由 n 个数据组成的待发布数据 $x_1,x_2,\cdots,x_n$，一致性约束条件表示为一个关于 x_i 的线性等式关系，如式(4)所示。

$$m_1x_1+m_2x_2+\cdots+m_nx_n=b \tag{4}$$

其中，m_j 和 b 是限定一致性约束条件的系数。根据上述定义，对于满足 m 个一致性约束条件差分隐私发布问题，在引入噪声机制之前，$\boldsymbol{x}$ 满足了如式(5)所示的一致性约束方程。

$$\boldsymbol{M}\boldsymbol{x}=\boldsymbol{b} \tag{5}$$

由式(5)可知，一致性约束问题取决于矩阵 $\boldsymbol{M}\in\mathbb{R}^{m\times n}$ 和向量 $\boldsymbol{b}\in\mathbb{R}^{m\times 1}$。由于添加噪声前的发布结果满足式(5)，因此式(5)为一致性方程，至少存在一个解。本文称满足式(5)的所有解均为一致性发布。记 $\mathcal{M}(D)$ 的输出为向量 $\tilde{\boldsymbol{x}}=\boldsymbol{x}+\boldsymbol{\xi}$，由上述分析可知，$\tilde{\boldsymbol{x}}$ 无法保证满足式(5)。为求得差分隐私下的最优一致性发布，文献[2,6,11-13]基于优化方程式(6)设计后置处理算法求得最优一致性发布 $\bar{\boldsymbol{x}}$。通常情况下，$\bar{\boldsymbol{x}}$ 的总体误差小于 $\tilde{\boldsymbol{x}}$，后置处理总是能有效地提升数据发布的精确性。

$$\begin{aligned}\min_{\bar{\boldsymbol{x}}}\quad & \|\bar{\boldsymbol{x}}-\tilde{\boldsymbol{x}}\|\\ \text{s.t.}\quad & \boldsymbol{M}\bar{\boldsymbol{x}}=\boldsymbol{b}\end{aligned} \tag{6}$$

根据一致性发布的存在性，定理 1 论证关于优化式(6)的最优一致性发布存在且唯一，同时最优一致性发布具有明确的解析表达式。

定理 1 优化方程式(6)存在求得最优一致性发布的解析表达式，记解析表达式为函数 $f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x})$，则 $\bar{\boldsymbol{x}}=f_{\boldsymbol{M},\boldsymbol{b}}(\tilde{\boldsymbol{x}})$。$f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x})$ 的表达式为

$$f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x})=\boldsymbol{x}+\boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{M}\boldsymbol{x}) \tag{7}$$

证明 令 $\boldsymbol{x}'=\bar{\boldsymbol{x}}-\tilde{\boldsymbol{x}}\Rightarrow\bar{\boldsymbol{x}}=\boldsymbol{x}'+\tilde{\boldsymbol{x}}$，则式(6)变换为

$$\begin{aligned}\min_{\boldsymbol{x}'}\quad & \|\boldsymbol{x}'\|\\ \text{s.t.}\quad & \boldsymbol{M}\boldsymbol{x}'=\boldsymbol{b}-\boldsymbol{M}\tilde{\boldsymbol{x}}\end{aligned}$$

该优化方程是关于 $\boldsymbol{x}'$ 的一致方程的最小范数解。由文献[14]可知

$$\boldsymbol{x}'=\boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{M}\tilde{\boldsymbol{x}})$$

即 $\bar{\boldsymbol{x}}=\tilde{\boldsymbol{x}}+\boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{M}\tilde{\boldsymbol{x}})$。因此

$$f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x})=\boldsymbol{x}+\boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{M}\boldsymbol{x})$$

证毕。

作为最优一致性发布的解析表达式，式(7)可作为通用解法有效解决任意差分隐私下最优一致性约束问题。然而，式(7)所涉及的 $\boldsymbol{M}^{\dagger}$ 是关于矩阵 $\boldsymbol{M}$ 的 Moore-Penrose 逆[14]运算。作为传统矩阵逆运算的拓展，Moore-Penrose 逆求解过程十分复杂，运算量极大。其求解难度不低于时间复杂度为 $O(n^3)$ 的传统求逆运算，无法高效地解决最优一致性约束发布问题。这导致通用解法难以有效解决数据规模较大的一致性约束问题。虽然通用解法的实用性有限，但作为小型最优一致性约束发布问题的解决方案仍然是合适的。

相较而言，针对具体发布问题设计的最优一致性发布算法的求解效率则高得多。Hay 等[2]设计的 Boosting 只需对完全 k 叉树分别执行一次自底向上和自顶向下的后置处理，即可实现最优一致性发布，时间复杂度仅为 $O(n)$；张双越等[7]设计的算法巧妙地利用轨迹流量发布问题的稀疏性实现多达

数十万个节点的交通路网的最优一致性发布。然而，上述两项技术并不具备通用性，难以适用于其他发布场景，甚至无法直接应用于拓展模型。因此，适用范围相对有限。

4　多重一致性约束问题的逼近方法

根据差分隐私多重一致性约束最优发布问题的思想，复杂的差分隐私的最优一致性约束问题可划分为多个最优一致性发布子问题。相比于原问题，合理划分后的子问题往往更简单且容易解决，或者可利用现有技术得以高效求解。由于文献[2-7]已对诸多子问题提供了解决方案，因此本文将重点研究如何利用各部分子问题的最优一致性发布结果实现原问题的最优一致性发布。构建差分隐私多重一致性约束最优发布问题首先进行子问题划分。由于 $\boldsymbol{M}$ 和 $\boldsymbol{b}$ 的每行代表了一个一致性约束条件，因此划分子问题的过程即将一致性约束条件重新排列、分组的过程。形式上相当于对 $\boldsymbol{M}$ 和 $\boldsymbol{b}$ 按行进行矩阵分块的过程，且表述同一个一致性约束子问题的所有一致性约束条件将被划分到同一个子矩阵。设原问题划分为 k 重一致性约束发布问题，则分块过程为

$$\boldsymbol{M}=\begin{bmatrix}\boldsymbol{M}_1\\ \boldsymbol{M}_2\\ \vdots\\ \boldsymbol{M}_k\end{bmatrix}$$

$$\boldsymbol{b}=\begin{bmatrix}\boldsymbol{b}_1\\ \boldsymbol{b}_2\\ \vdots\\ \boldsymbol{b}_k\end{bmatrix}$$

每个分块 $\boldsymbol{M}_i\in\mathbb{R}^{m_i\times n}$ 和 $\boldsymbol{b}_i\in\mathbb{R}^{m_i\times 1}$ 对应第 i 个最优一致性发布子问题，包含了 m_i 个一致性约束条件。由式(7)可知，对于噪声向量 $\tilde{\boldsymbol{x}}$，第 i 个子问题的最优一致性发布记为 $\bar{\boldsymbol{x}}=f_{\boldsymbol{M}_i,\boldsymbol{b}_i}(\tilde{\boldsymbol{x}})$。为简化表述，本文将 $f_{\boldsymbol{M}_i,\boldsymbol{b}_i}(\tilde{\boldsymbol{x}})$ 记为 $f_i(\tilde{\boldsymbol{x}})$。根据子问题的最优一致性发布，本文提出了差分隐私下多重一致性约束问题的逼近方法，该方法可被表述为极限表达式，如式(8)所示。

$$\bar{\boldsymbol{x}}=f_{\boldsymbol{M},\boldsymbol{b}}(\tilde{\boldsymbol{x}})=\lim_{t\to\infty}\left(f_k*\cdots*f_2*f_1\right)^t(\tilde{\boldsymbol{x}}) \tag{8}$$

其中，$*$ 为函数的复合运算符，即 $f_j*f_i(x)=f_j(f_i(x))$；t 表示函数的复合运算次数，即 $f^2(x)=f(f(x))$。根据极限表达式(8)，差分隐私下多重一致性约束问题的逼近方法的核心思想是通过依次反复求解一致性约束子问题，最终求解结果趋近于 $f_{\boldsymbol{M},\boldsymbol{b}}(\tilde{\boldsymbol{x}})$，即原问题的最优一致性发布。这样只需求得子问题的最优一致性发布，即可解决原问题的最优一致性发布。

5　最优一致性发布的性质分析

确保差分隐私下多重一致性约束问题的逼近方法可行的关键在于论证式(8)能否准确地收敛于原问题的最优一致性发布。由于该问题十分复杂，论证过程需要大量理论基础，本节首先从最优一致性发布的性质入手开展研究工作，然后循序渐进地寻找该问题的答案。

作为式(7)的关键组成部分，Moore-Penrose 逆 $\boldsymbol{M}^{\dagger}$ 具有一些重要的数学性质。相关资料[15-17]表明，$\boldsymbol{M}^{\dagger}$ 具有如下性质。

性质 1[15]　对于任意矩阵 $\boldsymbol{M}\in\mathbb{R}^{m\times n}$，都有 $\boldsymbol{M}^{\dagger}=\boldsymbol{M}^{\mathrm{T}}\left(\boldsymbol{M}\boldsymbol{M}^{\mathrm{T}}\right)^{\dagger}$ 成立。

性质 2[16]　对于任意矩阵 $\boldsymbol{M}\in\mathbb{R}^{m\times n}$，都有 $\boldsymbol{M}\boldsymbol{M}^{\dagger}\boldsymbol{M}=\boldsymbol{M}$ 成立。

性质 3　对于任意矩阵 $\boldsymbol{M}\in\mathbb{R}^{m\times n}$，都有 $\boldsymbol{M}^{\dagger}\boldsymbol{M}$ 为幂等矩阵成立，且有谱范数[17]满足 $0\leqslant\left\|\boldsymbol{M}^{\dagger}\boldsymbol{M}\right\|\leqslant 1$，$0\leqslant\left\|\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M}\right\|\leqslant 1$。

利用这些性质，本文通过进一步分析得出如下关于最优一致性发布的定理成立。

定理 2　对于任意向量 $\boldsymbol{x}\in\mathbb{R}^{n\times 1}$，设 $\boldsymbol{y}=f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x})$，则有 $\boldsymbol{M}\boldsymbol{y}=\boldsymbol{b}$。并且 $f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x})$ 的运算满足幂等律，即 $f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x})=f_{\boldsymbol{M},\boldsymbol{b}}*f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x})$。

证明　由于 $\boldsymbol{y}=f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x})$ 已为满足优化方程式(6)的最优一致性发布。将 $\boldsymbol{y}$ 代入式(6)中的 $\tilde{\boldsymbol{x}}$，显然 $\bar{\boldsymbol{x}}=\boldsymbol{y}$ 也是方程的一个可行解。此时，目标函数 $\left\|\bar{\boldsymbol{x}}-\tilde{\boldsymbol{x}}\right\|=0$，根据 $f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x})$ 的定义可知 $\boldsymbol{y}=f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{y})$。

设 $\boldsymbol{y}'=f_{\boldsymbol{M},\boldsymbol{b}}*f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x})=f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{y})=\boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{M}\boldsymbol{y})+\boldsymbol{y}$，由于 $\boldsymbol{M}\boldsymbol{y}=\boldsymbol{b}\Rightarrow\boldsymbol{b}-\boldsymbol{M}\boldsymbol{y}=\boldsymbol{0}$，代入可得 $\boldsymbol{y}'=\boldsymbol{y}$，因此幂等律得证。

根据定理 2，本文有如下推论。

推论 1　设 $\boldsymbol{p}\in\mathbb{R}^{n\times 1}$ 是任意满足 $\boldsymbol{M}\boldsymbol{p}=\boldsymbol{b}$ 的一致性发布，至少能够找到一个向量 $\boldsymbol{x}\in\mathbb{R}^{n\times 1}$ 使 $\boldsymbol{p}=f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x})$ 成立。

证明　根据定理 2 可知，任意满足 $\boldsymbol{M}\boldsymbol{p}=\boldsymbol{b}$ 的一

致性发布 $\boldsymbol{p}$ 都有 $\boldsymbol{p}=f_{M,b}(\boldsymbol{p})$。只需令 $\boldsymbol{x}=\boldsymbol{p}$，即找到一个向量 $\boldsymbol{x}$ 使 $\boldsymbol{p}=f_{M,b}(\boldsymbol{x})$ 成立。证毕。

虽然推论 1 只论证了 $\boldsymbol{p}$ 本身能满足推论条件，但实际上满足 $\boldsymbol{p}=f_{M,b}(\boldsymbol{x})$ 的向量往往无穷多，不过本文的分析过程只需关注其存在性，对具体有哪些 $\boldsymbol{x}$ 满足 $\boldsymbol{p}=f_{M,b}(\boldsymbol{x})$ 将不再赘述。

接下来，定理 3 将揭示最优一致性发布与其他一致性发布之间的关系。

定理 3 对于任意向量 $\boldsymbol{x}$ 及其最优一致性发布 $\boldsymbol{y}=f_{M,b}(\boldsymbol{x})$，设 $\boldsymbol{p}$ 是满足 $\boldsymbol{Mp}=\boldsymbol{b}$ 的一致性发布且 $\langle \boldsymbol{x}-\boldsymbol{p},\boldsymbol{y}-\boldsymbol{p}\rangle=0$，则 $\boldsymbol{p}$ 是关于 $\boldsymbol{x}$ 的最优一致性发布。

证明 采用反证法证明，若 $\boldsymbol{p}$ 不是关于 $\boldsymbol{x}$ 的最优一致性发布，即 $\boldsymbol{p}\neq\boldsymbol{y}$，则 $\|\boldsymbol{y}-\boldsymbol{p}\|>0$。

由于 $\boldsymbol{p}$ 是满足 $\boldsymbol{Mp}=\boldsymbol{b}$ 的一致性发布，根据推论 1，可令向量 $\boldsymbol{q}$ 使 $\boldsymbol{p}=f_{M,b}(\boldsymbol{q})$，有

$$\begin{aligned}&\|\boldsymbol{y}-\boldsymbol{p}\|=\|f_{M,b}(\boldsymbol{x})-f_{M,b}(\boldsymbol{q})\|=\\&\|\boldsymbol{x}-\boldsymbol{q}+\boldsymbol{M}^{\dagger}\boldsymbol{M}(\boldsymbol{x}-\boldsymbol{q})\|=\\&\|(\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})\|>0\end{aligned}$$

为求解 $\langle \boldsymbol{x}-\boldsymbol{p},\boldsymbol{y}-\boldsymbol{p}\rangle$，可对该内积的两部分分别展开。

$\boldsymbol{x}-\boldsymbol{p}$ 展开的结果为

$$\begin{aligned}&\boldsymbol{x}-\boldsymbol{p}=\boldsymbol{x}-\boldsymbol{q}-\boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{Mq})=\\&\boldsymbol{x}-\boldsymbol{q}-\boldsymbol{M}^{\dagger}\boldsymbol{b}+\boldsymbol{M}^{\dagger}\boldsymbol{Mq}=\\&\boldsymbol{x}-\boldsymbol{q}-\boldsymbol{M}^{\dagger}\boldsymbol{b}+\boldsymbol{M}^{\dagger}\boldsymbol{Mq}-\boldsymbol{M}^{\dagger}\boldsymbol{Mx}+\boldsymbol{M}^{\dagger}\boldsymbol{Mx}=\\&(\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})-\boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{Mx})\end{aligned}$$

$\boldsymbol{y}-\boldsymbol{p}$ 展开的结果为

$$\boldsymbol{y}-\boldsymbol{p}=(\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})$$

综上可得

$$\begin{aligned}\langle \boldsymbol{x}-\boldsymbol{p},\boldsymbol{y}-\boldsymbol{p}\rangle=&\|(\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})\|^2-\\&\langle \boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{Mx}),(\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})\rangle\end{aligned}$$

利用性质 1 可得

$$\begin{aligned}&\langle \boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{Mx}),(\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})\rangle=\\&\langle \boldsymbol{M}^{T}(\boldsymbol{MM}^{T})^{\dagger}(\boldsymbol{b}-\boldsymbol{Mx}),(\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})\rangle=\\&\langle (\boldsymbol{MM}^{T})^{\dagger}(\boldsymbol{b}-\boldsymbol{Mx}),\boldsymbol{M}(\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})\rangle\end{aligned}$$

利用性质 2 可得

$$\begin{aligned}&\langle (\boldsymbol{MM}^{T})^{\dagger}(\boldsymbol{b}-\boldsymbol{Mx}),\boldsymbol{M}(\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})\rangle=\\&\langle (\boldsymbol{MM}^{T})^{\dagger}(\boldsymbol{b}-\boldsymbol{Mx}),(\boldsymbol{M}-\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})\rangle=0\end{aligned}$$

因此，有

$$\langle \boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{Mx}),(\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})\rangle=0 \tag{9}$$

将式(9)代入 $\langle \boldsymbol{x}-\boldsymbol{p},\boldsymbol{y}-\boldsymbol{p}\rangle$ 的表达式，可得

$$\langle \boldsymbol{x}-\boldsymbol{p},\boldsymbol{y}-\boldsymbol{p}\rangle=\|(\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})\|^2>0$$

与题设不符，假设不成立。因此，$\boldsymbol{p}=\boldsymbol{y}$，$\boldsymbol{p}$ 是关于 $\boldsymbol{x}$ 最优一致性发布。证毕。

通常情况下，一致性发布的数量有无穷多个而最优一致性发布只有一个。定理 3 给出了判断某个一致性发布是否为最优一致性发布的方法，对于检验算法是否实现了最优一致性发布具有重要意义。

此外，研究还发现最优一致性发布满足 2 种不变性特征，分别是范数不变性以及内积不变性，具体内容如下。

定理 4 范数不变性。设 $\boldsymbol{p}$ 是满足 $\boldsymbol{Mp}=\boldsymbol{b}$ 的一致性发布，则对于向量 $\boldsymbol{x}$ 及其最优一致性发布 $\boldsymbol{y}=f_{M,b}(\boldsymbol{x})$，有

$$\|\boldsymbol{x}-\boldsymbol{p}\|^2=\|\boldsymbol{y}-\boldsymbol{x}\|^2+\|\boldsymbol{y}-\boldsymbol{p}\|^2 \tag{10}$$

证明 对 $\|\boldsymbol{x}-\boldsymbol{p}\|^2$ 展开，有

$$\begin{aligned}&\|\boldsymbol{x}-\boldsymbol{p}\|^2=\|(\boldsymbol{x}-\boldsymbol{y})+(\boldsymbol{y}-\boldsymbol{p})\|^2=\\&\|\boldsymbol{y}-\boldsymbol{x}\|^2+\|\boldsymbol{y}-\boldsymbol{p}\|^2+2\langle \boldsymbol{x}-\boldsymbol{y},\boldsymbol{y}-\boldsymbol{p}\rangle\end{aligned} \tag{11}$$

接下来，只需证明 $\langle \boldsymbol{x}-\boldsymbol{y},\boldsymbol{y}-\boldsymbol{p}\rangle=0$，定理 4 即可得证。根据推论 1，设向量 $\boldsymbol{q}$ 满足 $\boldsymbol{p}=f_{M,b}(\boldsymbol{q})$，则

$$\langle \boldsymbol{x}-\boldsymbol{y},\boldsymbol{y}-\boldsymbol{p}\rangle=-\langle \boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{Mx}),(\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})\rangle$$

由式(9)可知，$\langle \boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{Mx}),(\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{x}-\boldsymbol{q})\rangle=0$，则 $\langle \boldsymbol{x}-\boldsymbol{y},\boldsymbol{y}-\boldsymbol{p}\rangle=0$，代入式(11)可得式(10)成立。证毕。

定理 5 内积不变性。设 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$ 是满足方程 $\boldsymbol{Mp}=\boldsymbol{b}$ 的 2 个一致性发布，对于向量 $\boldsymbol{x}$ 及其最优一致性发布 $\boldsymbol{y}=f_{M,b}(\boldsymbol{x})$，则关于它们的内积满足

$$\langle \boldsymbol{p}_1-\boldsymbol{p}_2,\boldsymbol{x}-\boldsymbol{p}_2\rangle=\langle \boldsymbol{p}_1-\boldsymbol{p}_2,\boldsymbol{y}-\boldsymbol{p}_2\rangle \tag{12}$$

证明 由于 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$ 是满足方程 $\boldsymbol{Mp}=\boldsymbol{b}$ 的一致性发布，根据推论 1，可找到向量 $\boldsymbol{q}_1$ 和 $\boldsymbol{q}_2$ 满足 $\boldsymbol{p}_1=f_{M,b}(\boldsymbol{q}_1)$ 和 $\boldsymbol{p}_2=f_{M,b}(\boldsymbol{q}_2)$。则

$$\begin{aligned}&\langle \boldsymbol{p}_1-\boldsymbol{p}_2,\boldsymbol{x}-\boldsymbol{p}_2\rangle=\langle \boldsymbol{p}_1-\boldsymbol{p}_2,\boldsymbol{x}-\boldsymbol{y}+\boldsymbol{y}-\boldsymbol{p}_2\rangle=\\&\langle \boldsymbol{p}_1-\boldsymbol{p}_2,\boldsymbol{x}-\boldsymbol{y}\rangle+\langle \boldsymbol{p}_1-\boldsymbol{p}_2,\boldsymbol{y}-\boldsymbol{p}_2\rangle=\\&-\langle (\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{q}_1-\boldsymbol{q}_2),\boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{Mx})\rangle+\\&\langle \boldsymbol{p}_1-\boldsymbol{p}_2,\boldsymbol{y}-\boldsymbol{p}_2\rangle\end{aligned}$$

再次利用式(9)可知

$$\left\langle (\boldsymbol{I}-\boldsymbol{M}^{\dagger}\boldsymbol{M})(\boldsymbol{q}_1-\boldsymbol{q}_2), \boldsymbol{M}^{\dagger}(\boldsymbol{b}-\boldsymbol{M}\boldsymbol{x})\right\rangle = 0$$

同理，代入可得式(12)成立。证毕。

定理 4 和定理 5 分别体现了多重一致性约束问题的逼近方法迭代过程中内在的 2 种不变性特征，对于其收敛性的分析过程具有重大意义。

6　收敛性分析

根据上述分析结果，本节将进一步分析差分隐私下多重一致性约束问题的逼近方法的收敛性，并以此论证逼近方法经过多次迭代后将实现原问题的最优一致性发布。为了确保分析收敛性的过程便于理解，本节将依次从差分隐私下多重一致性约束问题的逼近方法能否收敛、收敛结果是否满足一致性约束以及一致性发布结果是否满足最优发布这3个问题逐步深入地进行收敛性分析。

首先是关于多重一致性约束问题的逼近方法能否收敛的分析。根据式(8)所示的计算过程，记第 s 次执行复合运算所得结果为 $\boldsymbol{x}_s$，$\boldsymbol{x}_0$ 表示执行一致性发布前的发布，即 $\boldsymbol{x}_0=\tilde{\boldsymbol{x}}$。根据定义，式(8)的复合函数计算过程实际上是一种自右向左的操作过程，记第 s 次计算过程所执行的函数为 $f_{[s]}(\boldsymbol{x})$，即对第 $[s]$ 个子问题求最优一致性发布，则 $[s]=(s-1)\bmod k+1$。设 $\boldsymbol{p}$ 为满足 $\boldsymbol{M}\boldsymbol{p}=\boldsymbol{b}$ 的任意一致性发布。由根据定理 4 所述的范数不变性，有

$$\|\boldsymbol{x}_i-\boldsymbol{p}\|^2=\|\boldsymbol{x}_{i+1}-\boldsymbol{x}_i\|^2+\|\boldsymbol{x}_{i+1}-\boldsymbol{p}\|^2$$

反复运用该定理，可得对于任意 s 有

$$\|\boldsymbol{x}_0-\boldsymbol{p}\|^2=\|\boldsymbol{x}_s-\boldsymbol{p}\|^2+\sum_{i=1}^{s}\|\boldsymbol{x}_i-\boldsymbol{x}_{i-1}\|^2$$

由于 $\|\boldsymbol{x}_0-\boldsymbol{p}\|^2$ 的值是有限且根据范数性质，级数 $\sum_{i=1}^{s}\|\boldsymbol{x}_i-\boldsymbol{x}_{i-1}\|^2$ 中各项非负，该级数为非降序列。由下列不等式可知，该级数收敛。

$$\lim_{s\to\infty}\sum_{i=1}^{s}\|\boldsymbol{x}_i-\boldsymbol{x}_{i-1}\|^2\leqslant\|\boldsymbol{x}_0-\boldsymbol{p}\|^2<\infty$$

由此可知，$\lim_{k\to\infty}\|\boldsymbol{x}_k-\boldsymbol{x}_{k-1}\|=0$，即序列 $\{\boldsymbol{x}_i\}$ 将收敛。因此，差分隐私下多重一致性约束问题的逼近方法随着迭代的进行必将收敛于某个发布结果。记迭代收敛于 $\boldsymbol{y}=\lim_{k\to\infty}\boldsymbol{x}_k$，则存在 $\boldsymbol{y}$ 使

$$\boldsymbol{y}=\lim_{t\to\infty}\left(f_k*\cdots*f_2*f_1\right)^t(\tilde{\boldsymbol{x}})$$

但是，上述结果无法确定关于 $\boldsymbol{y}$ 是否满足一致性发布。接下来，本文将尝试论证 $\boldsymbol{y}$ 是否满足方程 $\boldsymbol{M}\boldsymbol{y}=\boldsymbol{b}$ 的一致性发布。采用反证法论证，首先假设 $\boldsymbol{y}$ 不是满足 $\boldsymbol{M}\boldsymbol{y}=\boldsymbol{b}$ 的一致性发布，即 $\boldsymbol{M}\boldsymbol{y}\neq\boldsymbol{b}$。根据多重一致性约束问题的定义，必然存在某个 j 使 $\boldsymbol{M}_j\boldsymbol{y}\neq\boldsymbol{b}_j$。

令 $\boldsymbol{y}'$ 为原问题的一致性发布，即 $\boldsymbol{y}'=f_{M,b}(\tilde{\boldsymbol{x}})$，由定理 2 可知，$\boldsymbol{y}'$ 为 $\boldsymbol{M}_j\boldsymbol{y}=\boldsymbol{b}_j$ 的解。由于 $\boldsymbol{y}$ 不是 $\boldsymbol{M}_j\boldsymbol{y}\neq\boldsymbol{b}_j$ 的解，显然 $\boldsymbol{y}'$ 和 $\boldsymbol{y}$ 不同。令 $d=\|\boldsymbol{y}'-\boldsymbol{y}\|=\left\|\boldsymbol{M}_t^{\dagger}\left(\boldsymbol{b}_t-\boldsymbol{M}_t\boldsymbol{y}\right)\right\|$，有 $d>0$。

根据序列 $\{\boldsymbol{x}_i\}$ 的收敛性可知，对于任意 $\mu>0$，均存在足够大的数 l，可取任意 $s>l$，均有 $\|\boldsymbol{x}_{s-1}-\boldsymbol{y}\|<\mu$。此时，取任意满足 $s>l$ 且 $[s]=j$ 的整数 s，有 $\|\boldsymbol{x}_{s-1}-\boldsymbol{y}\|<\mu$。

然后，根据 $\|\boldsymbol{x}_{s-1}-\boldsymbol{y}\|$ 的分析，可得 $\|\boldsymbol{x}_{s-1}-\boldsymbol{y}\|$ 与 $\|\boldsymbol{x}_s-\boldsymbol{y}\|$ 满足如下关系。

$$\begin{aligned}
&\|\boldsymbol{x}_s-\boldsymbol{y}\|=\left\|\boldsymbol{M}_j^{\dagger}\left(\boldsymbol{b}_j-\boldsymbol{M}_j\boldsymbol{x}_{s-1}\right)+\boldsymbol{x}_{s-1}-\boldsymbol{y}\right\|=\\
&\left\|\boldsymbol{M}_j^{\dagger}\boldsymbol{b}_j-\boldsymbol{M}_j^{\dagger}\boldsymbol{M}_j\boldsymbol{x}_{s-1}+\boldsymbol{M}_j^{\dagger}\boldsymbol{M}_j\boldsymbol{y}-\boldsymbol{M}_j^{\dagger}\boldsymbol{M}_j\boldsymbol{y}+\boldsymbol{x}_{s-1}-\boldsymbol{y}\right\|=\\
&\left\|\boldsymbol{M}_j^{\dagger}\left(\boldsymbol{b}_j-\boldsymbol{M}_j\boldsymbol{y}\right)+\left(\boldsymbol{I}-\boldsymbol{M}_j^{\dagger}\boldsymbol{M}_j\right)\left(\boldsymbol{x}_{s-1}-\boldsymbol{y}\right)\right\|>\\
&\left\|\boldsymbol{M}_j^{\dagger}\left(\boldsymbol{b}_j-\boldsymbol{M}_j\boldsymbol{y}\right)\right\|-\left\|\left(\boldsymbol{I}-\boldsymbol{M}_j^{\dagger}\boldsymbol{M}_j\right)\left(\boldsymbol{x}_{s-1}-\boldsymbol{y}\right)\right\|=\\
&d-\left\|\left(\boldsymbol{I}-\boldsymbol{M}_j^{\dagger}\boldsymbol{M}_j\right)\left(\boldsymbol{x}_{s-1}-\boldsymbol{y}\right)\right\|
\end{aligned}$$

根据性质 3 可知，$0\leqslant\left\|\left(\boldsymbol{I}-\boldsymbol{M}_j^{\dagger}\boldsymbol{M}_j\right)\right\|\leqslant 1$，根据谱范数的性质有 $\left\|\left(\boldsymbol{I}-\boldsymbol{M}_j^{\dagger}\boldsymbol{M}_j\right)\left(\boldsymbol{x}_{s-1}-\boldsymbol{y}\right)\right\|\leqslant\|\boldsymbol{x}_{s-1}-\boldsymbol{y}\|$。因此，有

$$\begin{aligned}
&\|\boldsymbol{x}_s-\boldsymbol{y}\|=d-\left\|\left(\boldsymbol{I}-\boldsymbol{M}_j^{\dagger}\boldsymbol{M}_j\right)\left(\boldsymbol{x}_{s-1}-\boldsymbol{y}\right)\right\|\geqslant\\
&d-\|\boldsymbol{x}_{s-1}-\boldsymbol{y}\|>d-\mu
\end{aligned}$$

根据 $\|\boldsymbol{x}_s-\boldsymbol{y}\|>d-\mu$，只需取 $\mu<\frac{d}{2}$，就可得到 $\|\boldsymbol{x}_s-\boldsymbol{y}\|>d-\mu>d-\frac{d}{2}=\frac{d}{2}>\mu$。这显然与上述 $\{\boldsymbol{x}_i\}$ 的收敛性质相矛盾，假设不成立，$\boldsymbol{y}$ 是 $\boldsymbol{M}\boldsymbol{y}=\boldsymbol{b}$ 的解。因此，$\boldsymbol{y}$ 是满足方程 $\boldsymbol{M}\boldsymbol{y}=\boldsymbol{b}$ 的一致性发布。

最后，本文将进一步论证 $\boldsymbol{y}$ 不仅是满 $\boldsymbol{M}\boldsymbol{y}=\boldsymbol{b}$ 的一致性发布，而且 $\boldsymbol{y}$ 是关于 $\tilde{\boldsymbol{x}}$ 的最优一致性发布。

根据定理 5 所述的内积不变性，对于满足 $\boldsymbol{M}\boldsymbol{p}=\boldsymbol{b}$ 中的任意 2 个一致性发布 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$，有

$$\left\langle \boldsymbol{p}_1-\boldsymbol{p}_2, \boldsymbol{x}_{k-1}-\boldsymbol{p}_2\right\rangle=\left\langle \boldsymbol{p}_1-\boldsymbol{p}_2, \boldsymbol{x}_k-\boldsymbol{p}_2\right\rangle$$

反复运用该定理，可得

$$\langle \boldsymbol{p}_1-\boldsymbol{p}_2, \boldsymbol{x}_0-\boldsymbol{p}_2\rangle = \lim_{k\to\infty}\langle \boldsymbol{p}_1-\boldsymbol{p}_2, \boldsymbol{x}_k-\boldsymbol{p}_2\rangle = \langle \boldsymbol{p}_1-\boldsymbol{p}_2, \boldsymbol{y}-\boldsymbol{p}_2\rangle$$

由于 $\boldsymbol{y}$ 满足方程 $\boldsymbol{My}=\boldsymbol{b}$。不妨令 $\boldsymbol{p}_2=\boldsymbol{y}$，$\boldsymbol{p}_1=f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x}_0)$，代入可得

$$\langle f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x}_0)-\boldsymbol{y}, \boldsymbol{x}_0-\boldsymbol{y}\rangle = \langle f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x}_0)-\boldsymbol{p}_2, \boldsymbol{y}-\boldsymbol{y}\rangle = 0$$

结合定理 3，根据 $\langle f_{\boldsymbol{M},\boldsymbol{b}}(\boldsymbol{x}_0)-\boldsymbol{y}, \boldsymbol{x}_0-\boldsymbol{y}\rangle = 0$ 可知 $\boldsymbol{y}$ 是关于 $\tilde{\boldsymbol{x}}(\boldsymbol{x}_0)$最优一致性发布。因此，迭代的收敛结果 $\boldsymbol{y}$ 是关于 $\tilde{\boldsymbol{x}}$ 的最优一致性发布。

根据上述论证过程，本文成功证明差分隐私下多重一致性约束问题的逼近方法将会收敛于原问题的最优一致性发布。并且，逼近方法的收敛性是无条件的。即无论最优一致性子问题如何划分，逼近方法总能够成功实现最优一致性发布，体现了其强大的稳健性。因此，实践中只需考虑所划分子问题的易解性，通过合理的子问题划分提升发布效率，而不必担心划分结果能否正确实现最优一致性发布。

7 多重一致性约束问题的并行计算

由于差分隐私下多重一致性约束问题的逼近方法在划分子问题时只需考虑划分的合理性，合理的划分可使同类子问题间所涉及的子数据集互不相交，使同类子问题过程满足独立性。利用这种独立性设计并行的计算过程能够进一步提升多重一致性约束问题的求解效率。

以餐馆销量直方图发布问题为例，5 个单品对应了 5 个直方图一致性约束子问题，套餐一致约束子问题数量与发布天数 T 一致。该问题是 $T+5$ 重一致性约束问题。不难发现，5 个直方图一致性约束子问题各自关联了一个单品，所涉及数据之间互无交集，最优一致性发布的求解结果也互不影响。因此，这 5 个直方图一致性约束子问题可并行计算。同理，套餐一致约束子问题关联的每日发布数据也互无交集，这些子问题的求解也具有可并行性。

根据上述分析，结合多重一致性约束问题的逼近方法，餐馆销量直方图发布问题可划分为 c（$c=5$）个直方图一致性约束子问题与 T 个套餐一致约束子问题两组。组内的各个子问题可并行地、独立地求解。

根据上述分析，本文将设计差分隐私下餐馆销量直方图发布问题的最优一致性发布并行求解算法。一方面，以顾客购买一次套餐作为事件提供事件级别差分隐私[18]保护，根据餐馆提供的套餐分析，顾客购买套餐最多可以拿到 5 份食物（套餐 1 和套餐 2），每日销售数据单独发布时数据敏感度为 5。

另一方面，根据 Boosting 算法的理论，直方图发布的敏感度[2]取决于树高 $h=\lceil \log_k T\rceil+1$。因此，采用拉普拉斯作为噪声机制，餐馆销量直方图发布问题的全局敏感度[10]为 $\Delta_1=\mathrm{ch}$。分析套餐一致性约束问题，根据套餐内容，本文关于每日销量应满足一致性约束方程 $\boldsymbol{Bv}_t=\boldsymbol{0}$。$\boldsymbol{v}_t\in\mathbb{R}^{5\times 1}$ 表示第 t 天的销量，其第 i 个元素 $v_{t,i}$ 即为当天第 i 个单品的销量。经分析，$\boldsymbol{B}$ 的内容如式(13)所示。

$$\boldsymbol{B}=\begin{pmatrix} 1 & -5 & 3 & 4 & 0 \\ -1 & -1 & 1 & 0 & 2 \end{pmatrix} \tag{13}$$

由于套餐一致性约束子问题仅为数据规模为 5 的一致性发布问题，本文直接采用通用解法求解最优一致性发布。根据上述分析，本文提出算法 1 求解差分隐私下餐馆销量直方图发布问题的最优一致性发布。算法 1 表明，本文提出的多重一致性约束问题逼近理论不仅能够将更复杂的差分隐私一致性约束问题拆分成简单的子问题，而且可以利用子问题将的独立性实现并行求解算法，极大提升了算法求解性能。

算法 1 餐馆销量直方图一致性并行发布算法

输入 餐馆销量数据集 D，隐私预算 ε

输出 差分隐私直方图一致性发布树 $\overline{\mathcal{T}}_i$

1) 根据 D 统计第 t 天第 i 号单品的销量 $v_{t,i}$，初始化迭代次数 $s=0$

2) 对于第 i 号单品，根据 $v_{1,i}, v_{2,i}, \cdots, v_{T,i}$ 建立查询树 $\mathcal{T}_i$，树中第 j 个节点的值为 $x_{j,i}$。记 $\sigma: t\to j$ 表示第 t 日销量到 $\mathcal{T}_i$ 中叶子节点编号 j 的映射关系，则 $v_{t,i}$ 对应于 $x_{\sigma(t),i}$。$\mathcal{T}_i$ 中非叶子节点 $x_{j,i}$ 的值满足

$$x_{j,i}=\sum_{j'\in \mathrm{child}(j)} x_{j',i}$$

3) 添加拉普拉斯噪声[15]实现 ε – 差分隐私得到 $\widetilde{\mathcal{T}}_i$，满足

$$\tilde{x}_{j,i}^{(s)}=x_{j,i}+\xi, \xi\sim \mathrm{Lap}\left(\frac{\Delta_1}{\varepsilon}\right)$$

4) while 终止条件未达到

① 对于 $\mathcal{T}_i(1\leqslant i\leqslant c)$ 并行计算如下

采用 Boosting 算法对 $\tilde{x}_{1,i}^{(s)}, \tilde{x}_{2,i}^{(s)}, \cdots, \tilde{x}_{|T|,i}^{(s)}$ 求满足直

方图一致性约束的最优一致性发布 $\tilde{x}_{1,i}^{\prime(s)},\tilde{x}_{2,i}^{\prime(s)},\cdots,\tilde{x}_{|T|,i}^{\prime(s)}$

② 对于 $t(1\leqslant t\leqslant T)$ 日销量并行计算如下

结合 $\boldsymbol{B}$ 采用通用解法对 $\tilde{x}_{1,i}^{\prime(s)},\tilde{x}_{2,i}^{\prime(s)},\cdots,\tilde{x}_{|T|,i}^{\prime(s)}$ 求满足套餐一致性约束的最优一致性发布 $\tilde{x}_{\sigma(t),1}^{(s+1)},\tilde{x}_{\sigma(t),2}^{(s+1)},\cdots,\tilde{x}_{\sigma(t),c}^{(s+1)}$

③ 令 $s=s+1$

5) 令最优一致性发布 $\overline{T}_i$，树中 $\bar{x}_{t,i}=\tilde{x}_{t,i}^{(s)}$

6) return $\overline{T}_i$

8　实验分析

为了验证本文所提多重一致性约束问题逼近方法解决实际问题的效果，本文以餐馆销量直方图发布问题为例进行实验分析。实验将算法 1 与相应的通用解法对比，从算法求解效率、收敛性、稳定等方面对多重一致性约束问题的逼近方法进行综合分析。已有分析表明，Boosting 等差分隐私最优一致性发布问题的发布效果与加噪前数据内容无关[19]。为了实现更大规模的实验分析，在实验目的不受影响的前提下，本文采用虚拟数据进行实验。并且，为保持实验的准确与统一，实验均采用二叉树实现 Boosting 子算法。研究表明[4]，差分隐私一致性约束问题的最优一致性发布效果在不同 ε 下是稳定的，为避免实验冗长，实验统一设 $\varepsilon=1$。实验硬件环境如下：Intel®CoreTM i5-9500H CPU@3.00 GHz，8 GB 内存，1 TB 存储空间。

8.1　收敛性分析

作为一种逼近方法，算法 1 的收敛能力至关重要。因此，本节将通过跟踪算法运行过程来对算法的收敛性进行深入分析。收敛性分析包括 2 个方面，分别是一致性检验和发布误差分析。虽然通用结果可由解析表达式(7)直接求解，无法直接对比两者的迭代过程，但通用解法已被证明发布结果即为最优一致性发布，实验对比可作为检验算法 1 的一致性发布是否满足最优性的可靠标准。因此，在分析发布误差时，本文将其作为对比实验。由于通用解法需要消耗大量资源，常规的实验环境下通用解法难以有效满足发布天数远超 1 000 天的发布需求。因此，本实验以 $T=\{32,64,128,256,512,1024\}$ 天进行分组对比，实验迭代次数固定为 100 次，实验重复多次，记录平均结果。

作为多重一致性约束最优发布问题的基本目标，最终发布结果是否满足一致性约束是检验算法有效性的重要指标。为检验发布结果的一致性，本文提出了一致性偏差来衡量算法 1 在迭代过程中满足一致性的情况。将 s 次迭代后的所有数据 $\tilde{x}_{t,i}^{(s)}$ 组织为向量形式，记为 $\tilde{\boldsymbol{x}}^{(s)}$。然后，令 $\boldsymbol{\psi}^{(s)}=\boldsymbol{M}\tilde{\boldsymbol{x}}^{(s)}-\boldsymbol{b}$。根据一致性约束问题的定义可知，当 $\tilde{\boldsymbol{x}}^{(s)}$ 完全满足一致性时，$\boldsymbol{\psi}^{(s)}$ 应该等于 $\boldsymbol{0}$。不过，上述收敛性分析表明，逼近方法是在迭代过程中不断地令发布结果趋近于一致性。因而，本实验采用均方误差来衡量一致性偏差。记 s 次迭代后的一致性偏差为 mse_s，则 mse_s 的计算过程如式(14)所示。

$$\mathrm{mse}_s=\frac{1}{c|\mathcal{T}|}\sum_{t,i}\left(\psi_{t,i}^{(s)}\right)^2 \tag{14}$$

如图 1 所示，算法 1 在迭代过程中出现的一致性偏差随着迭代次数的增加而快速减少。虽然随着 T 的增加，一致性偏差的收敛速度有所减少，但所有实验都能在迭代 50 次左右使一致性偏差趋近于 0。因此，在迭代 50 次之后，算法就具备了较令人满意的一致性发布结果。并且随着迭代的增加，一致性偏差单调递减，表明算法 1 的发布结果具有较强稳定性，不会在迭代过程中突然出现不一致性变大的发布结果。

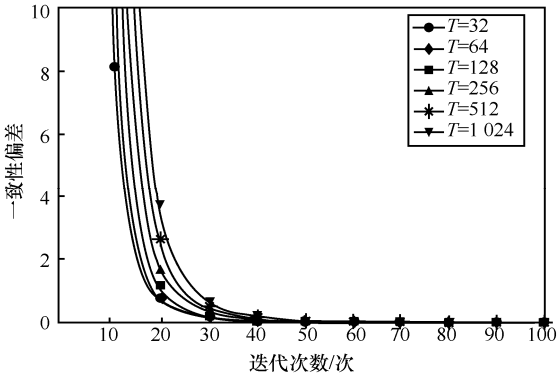


图 1　逼近方法的一致性偏差分析

除了发布结果的一致性，发布误差也是衡量发布结果优劣的重要指标。实验采用标准差衡量发布的误差。记 s 次迭代后发布结果相对于未加噪数据的标准差为 err_s，则 err_s 可由式(15)求得

$$\mathrm{err}_s=\sqrt{\frac{1}{c|\mathcal{T}|}\sum_{t,i}\left(\tilde{x}_{t,i}^{(s)}-x_{t,i}^{(s)}\right)^2} \tag{15}$$

图 2 中，虚线表示采用通用解法求得的最优一致性发布结果。由图 2 可以看出，无论发布天数 T

为多少，算法都能相对稳定地收敛于最优一致性发布对应的误差。并且在迭代前后，算法减少误差的效果十分明显，对于提升数据发布的精度具有重要价值。此外，从收敛效果来看，迭代初期算法即可平稳快速地收敛，使误差能够迅速逼近于最优一致性发布。算法只需要较少的迭代就能达到令人满意在一致性发布效果。因此，本文所提多重一致性约束问题的逼近方法具有较高的收敛能力以及算法稳定性。

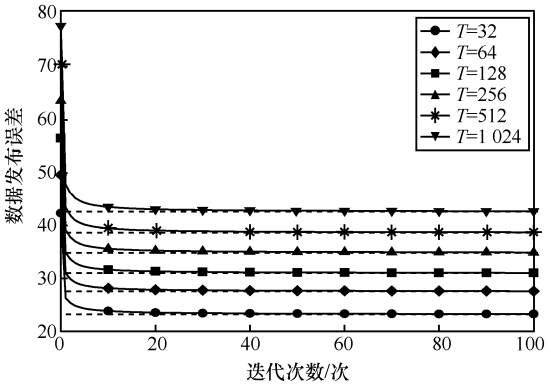


图 2 逼近方法的发布数据误差

8.2 求解效率分析

为进一步验证逼近方法的实用性，本节将探讨算法 1 求解最优一致性发布的效率。与 8.1 节实验不同，本次实验要求算法 1 达到足够的精度才停止。因此，实验设置算法终止条件为

$$\frac{1}{c|T|}\sum_{t,i}\left|\tilde{x}_{t,i}^{(s)}-x_{t,i}^{(s-1)}\right|<10^{-6}$$

将算法 1 的逼近方法与通用解法对比，求得在不同的发布天数 T 下的算法运行时间如图 3 所示。

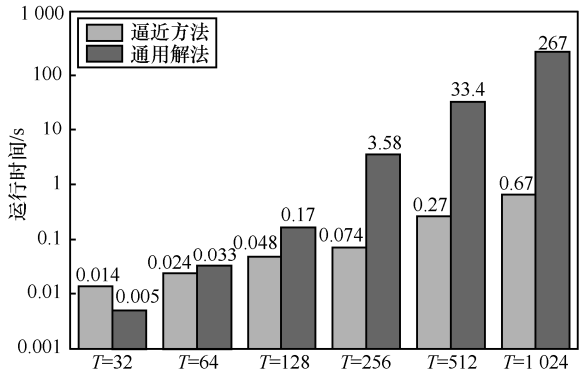


图 3 逼近方法与通用解法的运行时间对比

由图 3 可知，算法 1 的求解效率显著优于通用解法。从运行时间的增长幅度来看，算法 1 的运行时间随着数据量的增大接近于线性增长，与理论的时间复杂度 $O(Ts)$ 相符。而通用解法增长幅度则快很多，其时间复杂度为 $O(T^3)$。虽然当处理小规模数据时，算法 1 由于多次迭代运行时间略大于通用解法，但当数据规模变大时，通用解法的效率却低很多，仅处理 1 024 天的数据发布就需耗时多达 267 s，而算法 1 仅需要 0.667 s，差距高达 400 倍。

实际上，算法 1 所能处理的数据规模远不止 1 024 天。为探究其数据处理潜力，本文采用更大规模数据对其进行实验并记录求解耗时。实验结果如图 4 所示。图 4 表明，算法 1 已具备处理超大规模数据发布的能力，其所能处理的天数已高达百万。这表明算法 1 具有强大的数据处理能力，能够满足大多数实际发布的需要。同时也证明了本文所提出的多重一致性约束问题的逼近方法不仅具有较强的理论价值，还具有较强的实际应用价值。

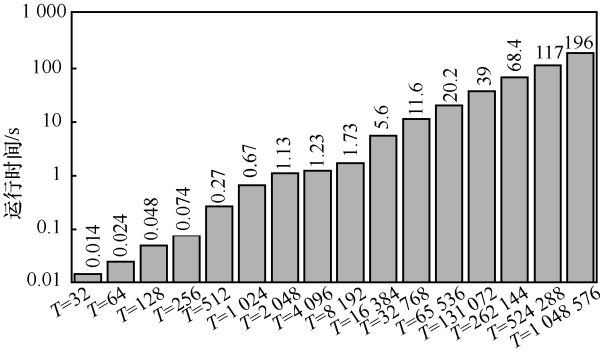


图 4 逼近方法在大规模数据下的运行时间

9 结束语

通过差分隐私下多重一致性约束问题的深入研究，本文提出并论证了多重一致性约束问题的逼近方法的有效性，为利用一致性约束子问题解决复杂的差分隐私一致性约束问题的方法奠定了扎实的理论基础。并且，本文以餐馆销量直方图发布问题为例设计的餐馆销量直方图一致性并行发布算法不仅充分展示了逼近方法较高的收敛能力以及求解效率，还体现了该方法具备的并行计算优势。研究结果表明，多重一致性约束问题的逼近方法具有较高的应用价值。

后续的研究工作中将以本文的研究成果作为理论基础，尝试将已被研究的差分隐私一致性发布模型推广到交通、医疗等领域，结合这些领域原本涉及的一致性发布过程实现应用范围更广、复杂程

度更高的差分隐私数据发布算法；同时，还将对多重一致性约束问题进行更加深入的理论研究，就如何更加合理地划分一致性约束子问题、如何提升逼近方法的收敛效率以及在不等式约束下如何实现多重一致性最优发布等问题开展研究工作，从而形成关于差分隐私最优一致性发布更加完善的理论体系。

参考文献：

[1] DWORK C. Differential privacy[C]//The 33rd International Conference on Automata, Languages and Programming - Volume Part II. Berlin: Springer, 2006: 1-12.

[2] HAY M, RASTOGI V, MIKLAU G, et al. Boosting the accuracy of differentially private histograms through consistency[J]. Proceedings of the VLDB Endowment, 2010, 3(1/2): 1021-1032.

[3] 吴英杰, 陈鸿, 王一蕾, 等. 面向任意区间树结构的差分隐私直方图发布算法[J]. 模式识别与人工智能, 2015, 28(12): 1084-1092.
WU Y J, CHEN H, WANG Y L, et al. A differentially private histogram publication algorithm for arbitrary range tree structure[J]. Pattern Recognition and Artificial Intelligence, 2015, 28(12): 1084-1092.

[4] 孙岚, 康健, 吴英杰, 等. 异方差加噪下差分隐私流数据发布一致性优化算法[J]. 清华大学学报(自然科学版), 2019, 59(3): 203-210.
SUN L, KANG J, WU Y J, et al. Consistency optimization algorithm for differential privacy streaming data publication with non-uniform private budgets[J]. Journal of Tsinghua University (Science and Technology), 2019, 59(3): 203-210.

[5] 贾俊杰, 陈慧, 马慧芳, 等. 差分隐私的查询一致性约束研究[J]. 计算机工程与科学, 2020, 42(1): 71-79.
JIA J J, CHEN H, MA H F, et al. Query consistency constraints of differential privacy[J]. Computer Engineering & Science, 2020, 42(1): 71-79.

[6] LEE J, WANG Y, KIFER D. Maximum likelihood postprocessing for differential privacy under consistency constraints[C]// The 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2015: 635-644.

[7] 张双越, 蔡剑平, 田丰, 等. 差分隐私下满足一致性的轨迹流量发布方法[J]. 计算机科学与探索, 2018, 12(12): 1903-1913.
ZHANG S Y, CAI J P, TIAN F, et al. Trajectory flow releasing method with consistency constraint under differential privacy[J]. Journal of Frontiers of Computer Science and Technology, 2018, 12(12): 1903-1913.

[8] CORMODE G, PROCOPIUC C, SRIVASTAVA D, et al. Differentially private spatial decompositions[J]. arXiv Preprint, arXiv:1103.5170, 2011.

[9] DWORK C, KENTHAPADI K, MCSHERRY F, et al. Our data, ourselves: privacy via distributed noise generation[C]//Advances in Cryptology - EUROCRYPT 2006. Berlin: Springer, 2006: 486-503.

[10] DWORK C, MCSHERRY F, NISSIM K, et al. Calibrating noise to sensitivity in private data analysis[C]//Theory of Cryptography Conference. Berlin: Springer, 2006: 265-284.

[11] LI C, HAY M, RASTOGI V, et al. Optimizing linear counting queries under differential privacy[C]// The 29th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems of Data. New York: ACM Press, 2010: 123-134.

[12] LEE J, CLIFTON C W. Top-k frequent itemsets via differentially private FP-trees[C]//The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Pree, 2014: 931-940.

[13] CORMODE G, PROCOPIUC C, SRIVASTAVA D, et al. Differentially private spatial decompositions[C]//2012 IEEE 28th International Conference on Data Engineering. Piscataway: IEEE Press, 2012: 20-31.

[14] DWYER P S, RAO C R, MITRA S K. Generalized inverse of matrices and its applications[J]. Journal of the American Statistical Association, 1973, 68(341): 239.

[15] GRAYBILL F A, MEYER C D, PAINTER R J. Note on the computation of the generalized inverse of a matrix[J]. SIAM Review, 1966, 8(4): 522-524.

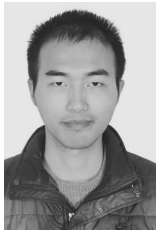
[16] PENROSE R. A generalized inverse for matrices[J]. Mathematical Proceedings of the Cambridge Philosophical Society, 1955, 51(3): 406-413.

[17] TÜRKMEN R, GÖKBAŞ H. On the spectral norm of r-circulant matrices with the Pell and Pell-Lucas numbers[J]. Journal of Inequalities and Applications, 2016, 2016(1): 1-7.

[18] CHEN Y, MACHANAVAJJHALA A, HAY M, et al. PeGaSus: data-adaptive differentially private stream processing[C]//The 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM Press, 2017: 1375-1388.

[19] QARDAJI W, YANG W N, LI N H. Understanding hierarchical methods for differentially private histograms[J]. Proceedings of the VLDB Endowment, 2013, 6(14): 1954-1965.

[作者简介]



蔡剑平（1990− ），男，福建漳州人，福州大学博士生，主要研究方向为差分隐私、矩阵分析及理论、最优化理论、大数据技术及理论等。

刘西蒙（1988− ），男，陕西西安人，博士，福州大学研究员、福建省“闽江学者”特聘教授，主要研究方向为隐私计算、密文数据挖掘、大数据隐私保护、可搜索加密等。

熊金波（1981− ），男，湖南益阳人，博士，福建师范大学教授，主要研究方向为安全深度学习、移动群智感知、隐私保护技术等。

应作斌（1982− ），男，安徽芜湖人，博士，新加坡南洋理工大学在站博士后，主要研究方向为基于属性的加密、区块链及隐私保护机器学习。

吴英杰（1979− ），男，福建泉州人，博士，福州大学教授、博士生导师，主要研究方向为数据安全隐私保护、工业大数据分析、智能医疗。