key-valueデータにおける局所差分プライバシーアルゴリズムPrivKVの改良

堀込光[1] 菊池 浩明[2] Chia-Mu Yu[3]

概要：局所差分プライバシは，単一次元の個人の持つプライバシー情報に局所的にノイズを付与することで，プライバシー情報が特定されることを防ぐ技術である．Randmized Response(RR) や Harmony のような従来の局所差分プライバシアルゴリズムでは，単一次元の情報しか扱うことができなかった．Ye らによって提案された局所差分プライバシアルゴリズム PrivKV では，離散値と連続値の 2 次元のデータである key-value データについて，離散値と連続値の相関を維持したプライバシー情報の収集を可能にした．しかし，PrivKV では，最尤推定法で集計がされており，精度が十分ではない．そこで，本稿では，PrivKV に対して Expectation Maximization(EM) アルゴリズムを適用する手法を提案し，数値実験により従来手法との精度を比較する．

Improvement of Local Differential Privacy algorithm PrivKV for key-value datasets

Hikaru Horigome[1] Hiroaki Kikuchi[2] Chia-Mu Yu[3]

1. はじめに

近年大幅に普及したスマートデバイスにより，サービス事業者は人々のあらゆる行動を分析できるようになった．例えば，Amazon などのオンライン商取引サービスでは，全利用者の購入履歴を取集し，購入頻度に基づいて利用者が購入した商品に関連する推薦商品を提供している．しかし，サービス事業者は全利用者の正確な行動履歴を保有しており，過失による情報漏洩や不正な内部犯行者によるプライバシ侵害の危険性がある．

個人情報の保護技術の一つに差分プライバシ [1] がある．これは、収集した情報の統計値を公開する際に確率的なノイズを付与するなどして出力プライバシーを保護する理論的な枠組みである．値を曖昧にする匿名加工情報よりも統計的な値の保護に適している．匿名加工情報は，再識別の禁止などの法的な規則と安全管理措置が適切である仮定の下で管理されており，プライバシ保護に関して理論的な保証はない．一方，収集の際の評価値などを保護する技術に局所差分プライバシ [2] がある．局所差分プライバシは，スマートデバイスからの情報を収集する際に確率的なノイズを付与するという技術である．これにより，サービス事業者でさえもユーザの真の値は分からない．

離散値と連続値の代表的な局所差分プライバシアルゴリズムとしてそれぞれ，Warner らによる Randomized Response[4] と Nguyên らによる Harmony[5] が知られている．さらに，Ye らは，Randomized Response と Harmony を組み合わせることで，離散値と連続値の組み合わせである key-value データセットに対して局所差分プライバシを満たす局所差分プライバシー方式 PrivKV[6] を提案した．これにより，各アイテムの頻度とその平均値を同時に推定することを可能にした．

PrivKV では頻度と平均値の推定に最尤推定法が用いられている．PrivKV における最尤推定法では，度数に負の値が生じる恐れがあり，度数が少ない key は誤差が大きくなる

[1] 明治大学 先端数理科学研究科
Guraduate School of Advanced Mathematical Sciences, Meiji University
[2] 明治大学 総合数理学部
School of Interdisciplinary Mathematical Science, Meiji University
[3] Department of Information Management and Finance, National Yang Ming Chiao Tung Univercity

問題がある．そこで，本研究では，PrivKV を用いて安全に収集したデータに対して，Expectation Maximization(EM) アルゴリズム [7] を用いて度数と平均値を推定する手法を提案する．逐次的にベイズ推定を繰り返すことで偏りのあるデータに対しても精度を高めた推定を期待できるためである．数値実験により，安全性指標 ϵ で最尤推定法を用いた PrivKV と比較した誤差の改善を試みる．

2. 準備

2.0.1 基本定義

各ユーザが自身のデータに対してノイズを付与し，そのデータを収集者へ送信する．収集者は，各ユーザから得られたデータを集計し，度数や平均値を推定する．ユーザ数を n とし，ユーザの集合を $U = \{u_1, u_2, ..., u_n\}$ とする．各ユーザは離散値，連続値，または key-value データを保持している．取扱う d 種類の離散値の集合を $K = \{k_1, k_2, ..., k_d\}$，$[-1, 1]$ の連続値の集合を V とする．プライバシー費用を ϵ とし，ある入力を t に対してランダムアルゴリズム M を適用することを $M(t, \epsilon)$ と記述する．

2.1 局所差分プライバシ

任意の異なる 2 つの入力に対して，M の出力が同一になる確率に差がないことを保証している．これにより，出力を参照してもユーザの正確な入力を特定することができず，ユーザのプライバシーを保護する．ランダムアルゴリズム M に対して局所差分プライバシは以下のように定義される．

定義 1.（局所差分プライバシ）

D を入力の集合，Z を出力の集合とする．M を入力 $t \in D$ に対して $z \in Z$ を出力するランダムアルゴリズムとする．任意の 2 つの入力 $t, t' \in D$ と任意の出力 $z \in Z$ に対して，

$$Pr[M(t, \epsilon) = z] \leq e^{\epsilon} Pr[M(t', \epsilon) = z]$$

が成立するとき，ランダムアルゴリズム M は ϵ-局所差分プライバシを満たすという．

2.2 Randomized Response(RR)

離散値データの局所差分プライバシアルゴリズムに Randomized Response(RR)[4] がある．RR では、確率 p で真の値を出力し，それ以外の確率 $q(= 1 - p)$ で偽の値を出力することで，プライバシを保護する．

入力 d 種類の値の集合 K 中からユーザが保有する値 $k \in K$ を入力とする．RR を用いて摂動化を行い，出力 $k^* \in K$ を収集者へ送信する．

摂動 確率 p で真の値 k を出力し，それ以外の確率 $q = (1 - p)$ で K の中から k 以外の値 $k' \in K - \{k\}$ を出力する．すなわち，

$$k^* = \begin{cases} k & w/p \quad p, \\ k' & w/p \quad q, \end{cases}$$

となる．

定理 1.（Randomized Response）[3]

維持確率 p，遷移確率 q が以下のとき，局所差分プライバシを満たす．

$$\begin{cases} p = \frac{e^{\epsilon}}{e^{\epsilon}+d-1}, \\ q = \frac{1}{e^{\epsilon}+d-1} \end{cases}$$

証明. 異なる 2 つの入力 $k, \hat{k} \in K$ に対して，同一の出力 k^* となるとき，入力に対する出力の確率比の最大は，

$$\begin{aligned} \frac{Pr[RR(k, \epsilon) = k^*]}{Pr[RR(\hat{k}, \epsilon) = k^*]} &\leq \frac{Pr[RR(k, \epsilon) = k]}{Pr[RR(\hat{k}, \epsilon) = k]} \\ &= \frac{p}{\frac{1-p}{d-1}} = \frac{p}{q} \\ &= e^{\epsilon} \end{aligned}$$

となる． □

集計 n 人のユーザから出力を収集し，各 $k_i \in K$ の度数を推定する．収集した出力の中で k_i の度数を f_i' とし，k_i の真の度数を f_i とする．最尤推定法では，k_i を保有する平均 $f_i p$ 人のユーザが k_i を出力し，k_i 以外を保有する平均 $(n - f_i)q$ 人のユーザが k_i を出力するため f_i' の期待値は，

$$f_i' = f_i p + (n - f_i)q$$

となる．上式から真の度数 f_i の最尤値は，

$$L[f_i] = \frac{f_i' - nq}{p - q} = \frac{n(p-1) + f_i'}{2p - 1}$$

となる．

2.3 Harmony

Nguyên らは連続データの局所差分プライバシアルゴリズムに Harmony[5] を提案している．Harmony では，連続値 $v \in V$ を 2 値化することで，RR の適用を可能にしている．また，RR を適用した値を定数倍することで，期待値を v として出力する．

入力 ユーザの保有する連続値 $v \in V(= [-1, 1])$ を入力とする．

摂動 Harmony の摂動工程を Value Perturbation Primitive(VPP) と呼ぶ．VPP には，連続値 v を 2 値化する工程と 2 値化された値 $v*$ に対して RR を適用する工程がある．

・2 値化 入力 v に対して，v に依存する確率で 2 値化された値を $v^*(\in \{-1, 1\})$ とする．

$$v^* = \begin{cases} 1 & w/p \quad \frac{1+v}{2}, \\ -1 & w/p \quad \frac{1-v}{2}. \end{cases}$$

・Randomized Response 2 値化された $v*$ に対して $RR(v*, \epsilon)$ を適用する．

$$v^+ = \begin{cases} v^* & w/p \quad p = \frac{e^\epsilon}{e^\epsilon+1}, \\ -v^* & w/p \quad q = \frac{1}{e^\epsilon+1} \end{cases}$$

Harmony では，VPP で得られた値に対して，定数倍した $\hat{v}$ を出力とする．

$$\hat{v} = v^+ \cdot \frac{e^\epsilon+1}{e^\epsilon-1}$$

定理 2.（VPP，Harmony）[5]

ランダムアルゴリズム VPP と Harmony は局所差分プライバシを満たす．

証明. 2 つの異なる入力 $v, v' \in V$ に対して同一の出力 $v^+ = 1$ となるとき，入力に対する出力の確率比は，

$$\begin{aligned} \frac{Pr[VPP(v,\epsilon)=1]}{Pr[VPP(v',\epsilon)=1]} &= \frac{\frac{1+v}{2}\cdot p + \frac{1-v}{2}\cdot q}{\frac{1+v'}{2}\cdot p + \frac{1-v'}{2} q} \\ &\leq \frac{\max_v\{v(e^\epsilon-1)+e^\epsilon+1\}}{\min_{v'}\{v'(e^\epsilon-1)+e^\epsilon+1\}} \\ &= e^\epsilon \end{aligned}$$

となる．$v^+ = -1$ についても同様に成立し，ϵ-局所差分プライバシを満たす．また，Harmony は VPP で得られた値を定数倍しているので，異なる 2 つの入力に対する出力の比は VPP と同様であり，ϵ-局所差分プライバシを満たす． □

このとき，入力 v に対する出力 $\hat{v}(=\{-\frac{e^\epsilon+1}{e^\epsilon-1}, \frac{e^\epsilon+1}{e^\epsilon-1}\})$ の期待値は，

$$\begin{aligned} E(\hat{v}) &= \frac{e^\epsilon+1}{e^\epsilon-1}(\frac{1+v}{2}p + \frac{1-v}{2}q) \\ &\quad -\frac{e^\epsilon+1}{e^\epsilon-1}(\frac{1+v}{2}q + \frac{1-v}{2}p) \\ &= \frac{e^\epsilon+1}{e^\epsilon-1}(vp - vq) \\ &= v\frac{e^\epsilon+1}{e^\epsilon-1}\frac{e^\epsilon-1}{e^\epsilon+1} \\ &= v \end{aligned}$$

となる．

集計 Harmony では，n のユーザから出力を収集し，平均値 m を推定する．収集した出力の中で，$\hat{v} = \frac{e^\epsilon+1}{e^\epsilon-1}$ の度数を m_1，$\hat{v} = -\frac{e^\epsilon+1}{e^\epsilon-1}$ の度数を $m_2(=1-m_1)$ とする．平均値 $\hat{m}$ は以下のように推定される．

$$\hat{m} = \frac{m_1 - m_2}{n}$$

2.4 PrivKV

Ye らは離散値と連続値の 2 次元データである key-value データについての局所差分プライバシアルゴリズム PrivKV を提案した [6]．key-value データの例は $\{<\text{YouTube}, 0.5>, <\text{Twitter}, 0.1>, <\text{Instagram}, 0.2>\}$ のような離散値 (アプリケーションなど) と連続値 (使用時間など) の組み合わせデータである．key-value セットの離散値に RR を，連続値に Harmony をそれぞれ独立に適用してしまうと離散値と連続値の相関が失われてしまう．そこで PrivKV では，入力が遷移するとき，離散値と連続値を同時に変化させることで，離散値と連続値の相関を維持した状態でデータ収集を行う．また，key に対する頻度と value に対する平均値を推定する．

i 番目のユーザ u_i が持つ l_i 個の key-value セットの集合を $S_i = \{<k_j, v_j> j1 \leq j \leq l_i, k_j \in K, v_j \in V\}$ とする．S_i を key-vlaue セット，S_i の h 番目の $<k_h, v_h>$ を key-value データと呼ぶ．

入力 d 種類の key-value データの収集を考える．$S'_i = \{<k'_s, v'_s> j1 \leq s \leq d, k'_s \in K, v'_s \in V\}$ とする．全ての $k' \in K$ について，key-value データが $<k_j, v_j> \in S_i$ の場合，$<k'_s, v'_s> = <1, v_j>$ とし，対応する key-value データが S_i にない場合，$<k'_s, v'_s> = <0, 0>$ とする．例えば，$d = 5$ で

$$S_i = \{<k_1, v_1>, <k_4, v_4>, <k_5, v_5>\}$$

であったとき，S'_i は，

$$S'_i = \{<1, v_1>, <0, 0>, <0, 0>, <1, v_4>, <1, v_5>\}$$

となり，このとき，$|S'_i| = d = 5$ である．こうして得られた S'_i を入力とする．摂動の工程には，value を摂動する工程と key を摂動する工程がある．

摂動 長さ d の key-value セット S'_i からランダムに 1 つの key-value データ $<k'_a, v'_a> \in S'_i$ を選択する．

- **value の摂動** v'_a の摂動には Harmony の摂動で用いられた VPP を使用する．$k'_a = 1$ の場合，ϵ_2 を用いて value に対して VPP を適用し，${v^+}_a = VPP(v'_a, \epsilon_2)$ とする．$k'_a = 0$ の場合，v'_a を $[-1, 1]$ からランダムに選択し，${v^+}_a = VPP(v'_a, \epsilon_2)$ とする．
- **key の摂動** key のランダマイズには，ϵ_1 を用いて $RR(k'_a, \epsilon_1)$ を使用する．PrivKV では，key が遷移するとき value も同時に変化させる．$k'_a = 1$ の場合，

$$<{k^*}_a, {v^+}_a> = \begin{cases} <1, {v^+}_a> & w/p \quad p = \frac{e^{\epsilon_1}}{1+e^{\epsilon_1}}, \\ <0, 0> & w/p \quad q = \frac{1}{1+e^{\epsilon_1}} \end{cases}$$

となり，$k'_a = 0$ の場合，

$$<{k^*}_a, {v^+}_a> = \begin{cases} <0, 0> & w/p \quad p = \frac{e^{\epsilon_1}}{1+e^{\epsilon_1}}, \\ <1, {v^+}_a> & w/p \quad q = \frac{1}{1+e^{\epsilon_1}} \end{cases}$$

となる．摂動化 $<{k^*}_a, {v^+}_a>$ と選択した key-value セットのインデックス a を送信する．

差分プライバシの合成定理により，差分プライバシを満たす複数のランダムアルゴリズムを多重に適用したアルゴリズムは全てのプライバシー費用 ϵ_i の和について差分プライバシを満たすことが知られている [9]．局所差分プライバシも同様であり，その際，全体のプライバシー費用 ϵ を求めることができる．

定理 3.（連続的な局所差分プライバシアルゴリズムの組み合わせによるプライバシ費用）

b 個のランダムアルゴリズムが累積した合成アルゴリズムを $\hat{M}$ とし，b 個のランダムアルゴリズムの集合を $M^* = \{M_1, M_2, ..., M_b\}$ とする．M^* のそれぞれのランダムアルゴリズム ${M^*}_i$ が ϵ_i-局所差分プライバシを満たすとき，b 個のランダムアルゴリズムが累積した合成アルゴリズム $\hat{M}$ は局所差分プライバシを満たし，全体の $\epsilon_{\hat{M}}$ は，

$$\epsilon_{\hat{M}} = \epsilon_{{M^*}_1} + \epsilon_{{M^*}_1} + \ldots + \epsilon_{{M^*}_b}$$

について $\epsilon_{\hat{M}}$-局所差分プライバシを満たす．

定理 1 の証明は，[9] を参照されたい．また，定理 1 より PrivKV 全体の ϵ は $\epsilon = \epsilon_1 + \epsilon_2$ となり，本稿では，$\epsilon_1 = \epsilon_2 = \frac{\epsilon}{2}$ として実験を行う．

集計　n 人ユーザからインデックス a と key-value セット $<{k^*}_a, {v^+}_a>$ を収集する．PrivKV では，$k_i \in K$ に対する頻度推定と $v_i \in V$ に対する平均値推定を目的とする．

- **頻度推定**　収集した key-value セット $<{k^*}_a, {v^+}_a>$ の中で，$k_i = 1$ の度数を f'_i とし，k_i の真の度数を f_i とする．RR 同様に最尤推定法では，k_i の度数の最尤値 $\hat{f}_i$ は，

$$\hat{f}_i = \frac{n(p-1) + f'_i}{2p-1}, \text{where } p = \frac{e^{\epsilon_1}}{1+e^{\epsilon_1}}$$

と推定される．

- **平均値推定**　収集した key-value セット $<{k^*}_a, {v^+}_a>$ の中で，$<k_i, v_i> = <1, 1>$ の度数を n'_{1i}，$<k_i, v_i> = <1, -1>$ の度数を n'_{2i} とする．最尤推定法を用いた $<k_i, v_i> = <1, 1>$ の推定度数 $\hat{n}_{1i}$，$<k_i, v_i> = <1, -1>$ の推定度数 $\hat{n}_{2i}$ は，

$$\hat{n}_{1i} = \frac{n(p-1) + n'_{1i}}{2p-1}$$
$$\hat{n}_{2i} = \frac{n(p-1) + n'_{2i}}{2p-1}, \text{where } p = \frac{e^{\epsilon_2}}{1+e^{\epsilon_2}}$$

となり，平均値 $\hat{m}_i$ は，

$$\hat{m}_i = \frac{\hat{n}_{1i} - \hat{n}_{2i}}{n}$$

と推定される．

2.4.1 PrivKVM

Ye らは，PrivKV の対話型アルゴリズム PrivKVM を提案している．摂動の工程で $<k'_a, v'_a> \in S'_i = <0, 0>$ が選択された場合，v'_a は [-1,1] からランダムに値が付与される．度数の少ない key では，v'_a がランダムに付与される割合が大きいため，平均値は 0 に近似する．PrivKVM では，算出した平均値をユーザに送り返すことで，この問題を改善している．

ユーザとの対話回数を $c(\geq 2)$ とし，c 回目の推定度数，推定平均値をそれぞれ $\hat{f}_i^{(c)}$，$\hat{m}_i^{(c)}$ とする．また，RR の工程と VPP の工程で対話ごとに割り振る ϵ をそれぞれ，$\{\epsilon_{11}, \epsilon_{12}, ..., \epsilon_{1c}\}$，$\{\epsilon_{21}, \epsilon_{22}, ..., \epsilon_{2c}\}$ とする．1 回目の収集では，PrivKV を用いて推定値 ${f_i}^{(1)}, {m_i}^{(1)} = PrivKV(S'_i, (\epsilon_{11} + \epsilon_{21}))$ を算出する．2 回目以降の収集では，$<k'_a, v'_a> \in S'_i = <0, 0>$ の場合，$v'_a = {m_i}^{(c-1)}$ とし，$VPP(<k'_a, v'_a>, \epsilon_{2c})$ と $RR(<k'_a, {v^*}_a>, \epsilon_{1c})$ を適用する．c 回の対話のあと，$\hat{f}_i^{(1)}$，$\hat{m}_i^{(c)}$ を推定値とする．また，$\epsilon_1 = \sum_c^{n=1} \epsilon_{1n}$，$\epsilon_2 = \sum_c^{n=1} \epsilon_{2n}$ となり，[6] では，

$$\begin{cases} \epsilon_{11} = \epsilon_1, \quad \epsilon_{12} = \epsilon_{13} = .. = \epsilon_{1c} = 0 \\ \epsilon_{21} = \epsilon_{22} = ... = \epsilon_{1c} = \frac{\epsilon_2}{c} \end{cases}$$

のように，ϵ を割り振っている．本稿でも同様の大きさで ϵ を割り振る．

3. 提案手法

上記 PrivKV の集計手法では，最尤推定法が用いられている．最尤推定法では，推定値が負の値を取る恐れがあることやデータの偏りがある場合，推定誤差が大きくなる問題がある．また，PrivKVM では，ユーザとの対話を行うため，対話コストがかかる．局所差分プライバシアルゴリズムにはベイズの定理を用いた反復手法を適用することで，推定精度が向上することが知られている．本稿では，key-value データにおける局所差分プライバシアルゴリズム PrivKV への EM アルゴリズムの適用手法を提案する．

3.1 EM アルゴリズム

EM アルゴリズムとは，ベイズの定理を利用した反復方式 [7] である．d 種類の入力の集合を $X = \{x_1, x_2 ..., x_d\}$，$d'$ 種類の出力の集合を $Z = \{z_1, z_2 ..., z_{d'}\}$ とする．n 人のユーザがそれぞれ自身の持つ値 $x_i \in X$ を入力とし，ランダムアルゴリズムを適用し，出力 $z_j \in Z$ を送信する．反復回数を t として，出力の集計から d 個の入力について度数推定を行う．t 回目の x_i 推定値を $(\theta^{(t)} = \{{\theta_1}^{(t)}, {\theta_2}^{(t)}, ..., {\theta_d}^{(t)})$ とし，初期値を $\theta^{(0)} = (\frac{1}{d}, \frac{1}{d}, ..., \frac{1}{d})$ とする．また，ユーザ u の t 回目の X に対する推定値を $\hat{\theta}_u^{(t)} = (\hat{\theta}_{u,1}^{(t)}, \hat{\theta}_{u,2}^{(t)}, ..., \hat{\theta}_{u,d}^{(t)})$ とする．

入力 x_i に対して出力 z_j となる条件付き確率は，

$$Pr[z_j|x_i] = \frac{Pr[z_j, x_i]}{Pr[x_i]}$$

となる．また，ベイズの定理より，出力 z_j で条件付けた入力が x_i である確率は，

$$Pr[x_i|z_j] = \frac{Pr[z_j|x_i]Pr[x_i]}{\sum_{s=1}^{|X|} Pr[z_j|x_s]Pr[x_s]}$$

となり，$t-1$ 回目の推定出力 z_j に対する入力 x_i の t 回目の推定確率は，

$$\hat{\theta}_{u,i}^{(t)} = Pr[x_i|z_j] = \frac{Pr[z_j|x_i]{\theta_i}^{(t-1)}}{\sum_{s=1}^{|X|} Pr[z_j|x_s]{\theta_s}^{(t-1)}}$$

で更新される．$t-1$ 回目の推定値 $\theta^{(t-1)}$ を用いて，ユーザごとに $\hat{\theta}_u^{(t-1)}$ を計算し，$\hat{\theta}_u^{(t-1)}$ の平均値を $\theta^{(t)}$ として更

新する．

$$\theta^{(t)} = \frac{1}{n}\sum_{u=1}^{n} {\theta_u}^{(t-1)}$$

これをあらかじめ定めた閾値 η>0 に対して，${\theta_i}^{(t)}$ が $|{\theta_i}^{(t)} - {\theta_i}^{(t-1)}| \leq \eta$ となって収束するまで繰り返す．これにより，入力 x_i の度数を推定する．

3.2 PrivKV への EM アルゴリズムの適用手法

ユーザ i の入力 S'_i について，摂動化する key-value データ $<k'_a, v'_a> \in S'_i$ の vlaue の値は $v'_a \in V(=[-1,1])$ であるため，出力 $<{k^*}_a, {v^+}_a>$ から $<k'_a, v'_a>$ を推定することは困難である．そこで，出力 $<{k^*}_a, {v^+}_a>$ から摂動工程の中の VPP を適用し value を 2 値化した key-value データ $<k'_a, {v^+}_a>$ の度数を EM アルゴリズムを用いて推定する．このとき，推定する度数の集合 X は，$X = \{<1,1>, <1,-1>, <0,1>, <0,-1>\}$ となり，出力の集合 Z は，$Z = \{<1,1>, <1,-1>, <0,0>\}$ となる．

n 人のユーザから出力 $<{k^*}_a, {v^+}_a> \in Z$ を観測する．$k_a \in K$ について初期値は $\theta^{(0)} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ とする．t 回の反復を行なった結果得られた入力 X の推定度数を $\theta^{(t)} = (\theta^{(t)}_{<1,1>}, \theta^{(t)}_{<1,-1>}, \theta^{(t)}_{<0,1>}, \theta^{(t)}_{<0,-1>})$ とする．k_a の度数 f_a は，

$$\hat{f}_a = n(\theta^{(t)}_{<1,1>} + \theta^{(t)}_{<1,-1>})$$

と推定でき，平均値 m_a は，

$$\hat{m}_a = \frac{(\theta^{(t)}_{<1,1>} - \theta^{(t)}_{<1,-1>})}{n}$$

と推定する．

証明. value を 2 値化した key-value データ $<{k^*}_a, {v^+}_a>$ が $<1,1>, <1,-1>$ であることは，k_a について v_a の値を保有していることを示めし，$<{k^*}_a, {v^+}_a>$ が $<0,1>, <0,-1>$ であることは k_a について v_a の値を保有していないことを示す．k_a について，ユーザ n の中で，$<{k^*}_a, {v^+}_a> = <1,1>$ の割合を $\delta_{<1,1>}$, $<{k^*}_a, {v^+}_a> = <1,-1>$ の割合を $\delta_{<1,-1>}$ とする．また，ユーザ n の入力 S'_i の中で，k_a についての値 v_a を保有しているユーザの割合を δ^* とする．$d(=|S'_i|)$ 種類の key からランダムに 1 つの key-value データ $<{k^*}_a, v_a>$ を選択するとすると，

$$E(\delta^*) = E(\delta_{<1,1>} + \delta_{<1,-1>})$$

となる．$\delta_{<1,1>}$ の推定値を $\theta_{<1,1>}$，$\delta_{<1,-1>}$ の推定値を $\theta_{<1,-1>}$ とすると，k_a について v_a の値を保有するユーザの割合の推定値は，$\theta_{<1,1>} + \theta_{<1,-1>}$ となるため，推定度数 $\hat{f}_a$ は，

$$\hat{f}_a = n(\theta_{<1,1>} + \theta_{<1,-1>})$$

となる．また，v_a の平均値 m_a は 2.3 節の VPP の期待値同様に，

$$E(m_a) = E(\frac{\delta_{<1,1>} - \delta_{<1,-1>}}{n})$$

となるので，推定平均値 $\hat{m}_a$ は

$$\hat{m}_a = \frac{(\theta^{(t)}_{<1,1>} - \theta^{(t)}_{<1,-1>})}{n}$$

となる． □

提案手法を図 1 に示す．

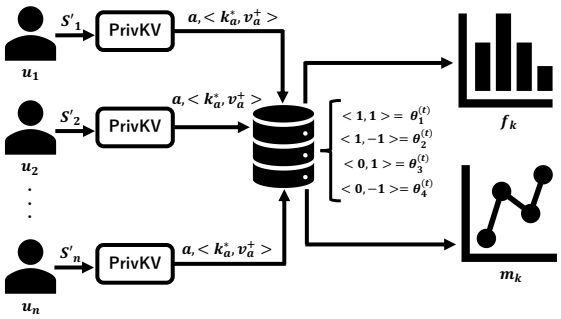

図 1 システム構成図

数値例 k_a の度数 $\hat{f}_a$ と平均値 $\hat{m}_a$ を推定することを考える．ユーザ u の出力 $<{k^*}_a, {v^*}_a> \in Z$ が $z_1 = <1,1>$ であったとする．度数を推定する key-value セット $X = \{<1,1>, <1,-1>, <0,1>, <0,-1>\}$ の度数初期値を $\theta^{(0)} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\})$ とする．出力 $<{k^*}_a, {v^*}_a>$ が $z_1 = <1,1>$ であったとき，value を 2 値化した key-value セット $<k'_a, {v^*}_a>$ が $x_1 = <1,1>$ である確率は，ベイズの定理を用いて，

$$\begin{aligned} Pr[x_1|z_1] &= \frac{Pr[z_1|x_1]Pr[x_1]}{\sum_{s=1}^{4} Pr[z_1|x_s]Pr[x_s]} \\ &= \frac{Pr[z_1|x_1]{\theta_1}^{(0)}}{\sum_{s=1}^{4} Pr[z_1|x_s]{\theta_s}^{(1)}} \\ &= \frac{\frac{1}{4}p_1p_2}{\frac{1}{4}p_1p_2 + \frac{1}{4}p_1q_2 + \frac{1}{4}q_1p_2 + \frac{1}{4}q_1q_2} \\ &= \frac{p_1p_2}{p_1(p_2+q_2) + q_1(p_2+q_2)} \\ &= p_1p_2 = \frac{e^{\epsilon_1}e^{\epsilon_2}}{(1+e^{\epsilon_1})(1+e^{\epsilon_2})} \end{aligned}$$

となる．$\epsilon = 1$，$\epsilon_1 = \epsilon_2 = \frac{\epsilon}{2}$ とすると，$\theta^{(1)}_{1,u}$ は，

$$\theta^{(1)}_{1,u} \approx 0.387455$$

となる．出力 z_1 に対する入力 x_2, x_3, x_4 の確率についても同様に計算し，全てのユーザの平均を $\theta^{(2)}$ として更新する．

4. 実験

ϵ の値とユーザ数 n を変化させ，PrivKV と同様の手法で収集した key-value セットに対して，EM アルゴリズムを用いて度数と平均値の推定を行い，最尤推定法を用いている

PrivKV や PrivKVM と推定精度を比較する．PrivKVM では推定平均値を用いてユーザと 3 回の対話を行い，EM アルゴリズムを用いた推定では対話を行わない．

4.0.1 データセット

データセットには，key と value がガウス分布 ($\mu = 0, \sigma = 10$)，べき分布 ($F(x) = (1+0.1x)^{-\frac{11}{10}}$)，線形分布 ($F(x) = x$) に従う 3 つの合成データを使用する．表 1 にユーザ数が 10^5 のときの合成データそれぞれの key に対する度数 f_i の平均，分散，value に対する平均値 m_i の平均，分散を示す．

表 1 ユーザ数 n=10^5，key 数 d=50 のデータセット

データ分布	$E(\frac{f_k}{n})$	$Var.(\frac{f_k}{n})$	$E(m_k)$	$Var.(m_k)$
ガウス分布	0.49506	0.10926	-0.00987	0.43702
べき分布	0.20660	0.062901	-0.58681	0.25160
線形分布	0.51	0.08330	0	0.34694

4.1 評価手法

n 人のユーザから出力された key-value データに対して，EM アルゴリズムと最尤推定法 (PrivKV，PrivKVM) で度数 $\hat{f}_k$ と平均値 $\hat{m}_k$ を推定する．key の真の度数を f_k，value の真の度数を m_k とし，推定誤差を MSE(Mean Square Error) を用いて以下のように評価する．

$$MSE_f = \frac{1}{|K|}\sum_{i=1}^{|K|}\left(\frac{\hat{f}_i}{n} - \frac{f_i}{n}\right)^2$$

$$MSE_m = \frac{1}{|K|}\sum_{i=1}^{|K|}(\hat{m}_i - m_i)^2$$

この試行を 10 回行い，評価値の平均を精度とする．

4.2 実験結果

4.2.1 ϵ による精度

ユーザ数 n=10^5，key 数 d=50 とし，ϵ を変化させ推定精度を比較する．表 2 にガウス分布，表 3 にべき分布，表 4 に線形分布における度数の推定誤差 MSE_f の平均値を示す．3 つの分布において，EM アルゴリズムを用いた手法では，最尤推定を用いた PrivKV と PrivKVM に比べ，どの ϵ の場合でも最も精度が高い結果となった．特に $\epsilon = 0.1$ のとき，線形分布では，EM アルゴリズによる推定度数は PrivKV に比べ 60.62% 改善されている．

また，図 2 にガウス分布，図 3 にべき分布，図 4 に線形分布における平均値の推定誤差 MSE_m の平均値を示す．平均値の推定誤差 MSE_m についてもどの ϵ の場合でも EM アルゴリズムを用いた手法の精度が最も高い．また，EM アルゴリズムを用いた手法では他手法に比べ，ϵ が大きくなるにつれより正確に推定される．図 5 に $\epsilon = 3$ のときの key に対する推定平均値の分布を示す．PrivKV では度数の割合が 0.3 以上の key のみを小さな誤差で推定しているのに対し，EM アルゴリズムを用いた手法では，度数の割合が 0.1 の key まで小さな誤差で推定していることがわかる．

表 2 ϵ による $MSE_f(\times 10^{-4})$ の変化 (ガウス分布)

ϵ	0.1	0.5	1	2
EM	756.67	63.939	18.075	4.636
PrivKV	1921.72	84.628	22.587	4.766
PrivKVM($c=3$)	1472.77	75.394	26.212	6.248
ϵ	3	4	5	
EM	2.018	1.626	1.147	
PrivKV	2.324	1.723	1.319	
PrivKVM($c=3$)	2.508	1.673	1.172	

表 3 ϵ による $MSE_f(\times 10^{-4})$ の変化 (べき分布)

ϵ	0.1	0.5	1	2
EM	671.25	55.478	18.578	4.876
PrivKV	2170.25	84.833	19.274	5.497
PrivKVM($c=3$)	851.21	62.403	23.182	5.254
ϵ	3	4	5	
EM	1.591	1.359	0.973	
PrivKV	2.587	1.394	1.019	
PrivKVM($c=3$)	2.420	1.365	0.992	

表 4 ϵ による $MSE_f(\times 10^{-4})$ の変化 (線形分布)

ϵ	0.1	0.5	1	2
EM	602.83	70.345	16.022	5.618
PrivKV	1885.28	92.988	20.173	7.404
PrivKVM($c=3$)	1462.74	82.794	18.440	6.157
ϵ	3	4	5	
EM	2.523	1.502	1.282	
PrivKV	2.790	1.943	1.428	
PrivKVM($c=3$)	2.597	1.913	1.279	

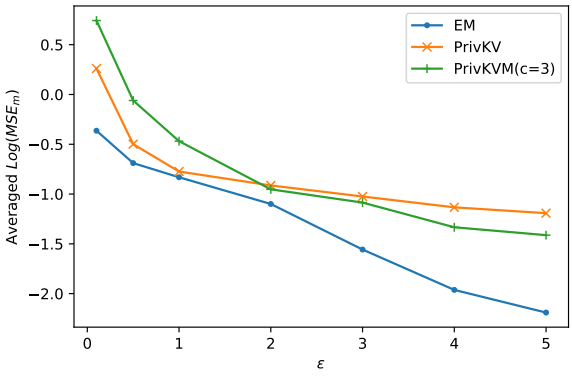

図 2 ϵ による MAE_m の変化 (ガウス分布)

4.2.2 ユーザ数による精度

次に $\epsilon = 2$，key 数 d=50 とし，ユーザ数を変化させ推定精度を比較する．表 5 にガウス分布，表 6 にべき分布，表 7 に線形分布における度数の推定誤差 MSE_f の平均値を

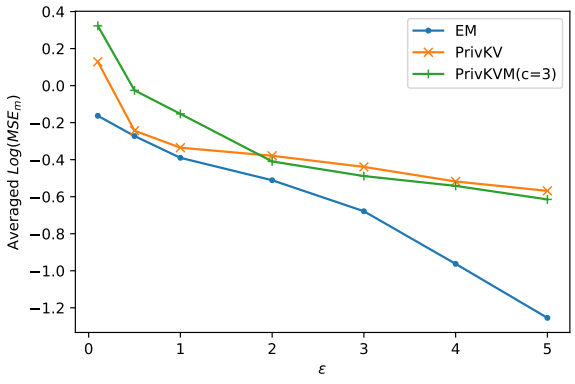

図 3 ϵ による MAE_m の変化 (べき分布)

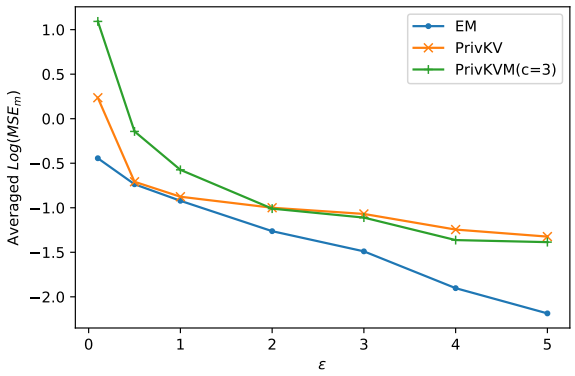

図 4 ϵ による MAE_m の変化 (線形分布)

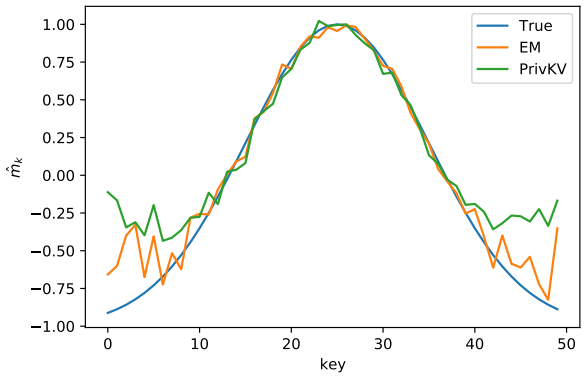

図 5 平均値推定 ($\epsilon = 3$)

示す．EM アルゴリズムを用いた手法では，ユーザ数に関係なく PrivKV，PrivKVM に比べ，推定誤差が小さかった．特に，べき分布では，$n = 1 \times 10^4$ のとき PrivKV に比べ精度が 36.2% 改善している．EM アルゴリズムを用いることで，度数の小さな key について，より正確に推定していると考えられる．

また，図 6 にガウス分布，図 7 にべき分布，図 8 に線形分布における度数の推定誤差 MSE_m の平均値を示す．ユーザ数の少ない $n \leq 5 \times 10^4$ では，EM アルゴリズムを

表 5 ユーザ数 $n(\times 10^4)$ による $MSE_f(\times 10^{-4})$ の変化 (ガウス分布)

$n(\times 10^4)$	1	5	10	30
EM	476.25	107.26	36.086	16.732
PrivKV	527.91	110.05	51.430	18.361
PrivKVM($c = 3$)	612.90	137.27	62.515	18.869
$n(\times 10^4)$	50	75	100	500
EM	9.414	6.519	4.635	1.823
PrivKV	11.552	7.182	4.765	2.473
PrivKVM($c = 3$)	12.242	7.489	6.248	2.034

表 6 ユーザ数 $n(\times 10^4)$ による $MSE_f(\times 10^{-4})$ の変化 (べき分布)

$n(\times 10^4)$	1	5	10	30
EM	346.71	72.822	39.234	13.459
PrivKV	543.28	99.441	51.115	17.552
PrivKVM($c = 3$)	424.72	95.245	62.975	15.104
$n(\times 10^4)$	50	75	100	500
EM	9.079	6.534	4.876	1.363
PrivKV	12.198	7.383	5.497	1.991
PrivKVM($c = 3$)	12.236	6.996	5.254	1.534

表 7 ユーザ数 $n(\times 10^4)$ による $MSE_f(\times 10^{-4})$ の変化 (線形分布)

$n(\times 10^4)$	1	5	10	30
EM	404.47	89.921	54.165	17.995
PrivKV	538.94	118.34	69.998	19.334
PrivKVM($c = 3$)	733.46	113.03	56.959	19.913
$n(\times 10^4)$	50	75	100	500
EM	10.482	6.549	5.618	1.802
PrivKV	15.932	7.801	7.404	2.003
PrivKVM($c = 3$)	15.133	8.607	6.157	1.906

用いた手法は他手法と比べ推定誤差の差は小さいが，ユーザ数が増えるにつれ推定誤差の差は大きくなり，より正確に推定している．特に度数の少ない key が多数存在するガウス分布では，$\varepsilon = 5$ のとき，31.6% の改善が見られた．これにより，ユーザ数に関わらず，対話を行う PrivKVM よりも高い精度で推定が可能であるため，対話コストを削減することができる．

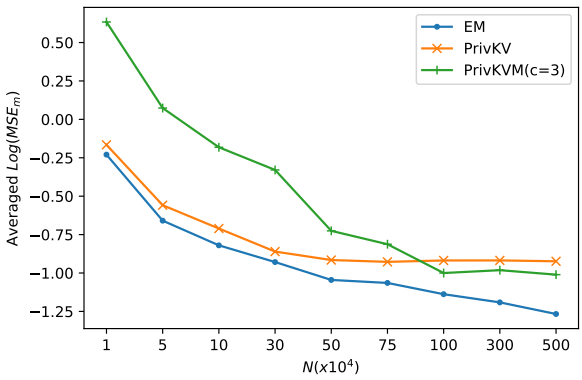

図 6 ユーザ数 n による MAE_m の変化 (ガウス分布)

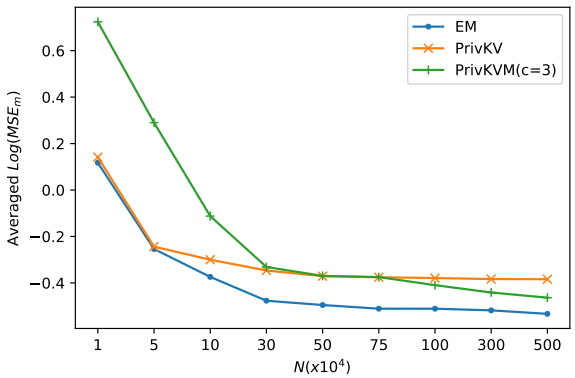

図 7 ユーザ数 n による MAE_m の変化 (べき分布)

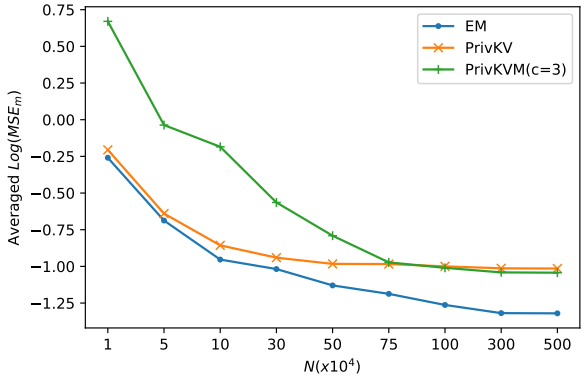

図 8 ユーザ数 n による MAE_m の変化 (線形分布)

5. おわりに

本稿では，key-value データにおける局所差分プライバシアルゴリズム PrivKV の推定精度を改善するため，PrivKV と同様の手法で収集したデータに対して EM アルゴリズムを適用する手法を提案した．合成データを用いて評価を行い，特に PrivKV，PrivKVM では大きな誤差となった度数の小さな key について，EM アルゴリズムを用いることで，推定精度が改善した．度数推定では，ユーザ数 $n = 10^4$，$\epsilon = 0.1$ のとき，3 つの合成データの平均で約 69.5% の改善が見られた．また，平均値推定では，ユーザ数 $n = 10^4$，$\epsilon = 5$ のとき，平均で 85.2% の改善が見られた．特に度数の小さな key が比較的多いべき分布に従うデータの時に EM アルゴリズムを用いた手法が効果的であった．また，商品の評価データのような度数の少ない key が多数存在するビックデータにおいても PrivKV に比べ精度よく推定できると考える．しかし，EM アルゴリズムを用いた場合でも，度数の小さな key ほど誤差が大きかった．度数の小さな key を正確に推定するアルゴリズムの作成を今後の課題とする．

参考文献

[1] C. Dwark, F. McSherry, K. Nissim, A. Smith, “Calibrating noise to sensitivity in private data analysis”, TCC, Vol. 3876, p. 265-284, 2006.
[2] J. C. Duchi, M. I. Jordan, M. J. Wainwright, “Local privacy and statistical minimax rates”, FOCS, pp. 429-438, 2013.
[3] P. Kairouz, S. Oh, and P. Viswanat, “Extremal mechanisms for local differential privacy”, NIPS, pp. 2879-2887, 2014.
[4] S. L. Warner, “Randomized response: A survey technique for eliminating evasive answer bias”, Journal of the American Statistical Association, pp. 63-69, 1965..
[5] T. T. Nguyên, X. Xiao, Y. Yang, S. C. Hui, H. Shin, J. Shin, “Collectiong and analyzing data from smart device users with local differential privacy”, arXiv:1606.05053, 2016.
[6] Q. Ye, H. Hu, X. Meng, H. Zheng, “PrivKV : Key-Value Data Collection with Local Differential Privacy”, IEEE S&P, pp. 294-308, 2019.
[7] 宮川雅巳, “EM アルゴリズムとその周辺”, 応用統計学, Vol. 16, No. 1, pp. 1-19, 1987.
[8] Ú. Erlingsson, V. Pihur, A. Korolova, “RAPPOR : Randomized Aggregatable Privacy-Preserving Ordinal Response”, ACM, pp. 1054-1067, 2014.
[9] F. McSherry, “Privacy integrated queries: an eztensible platform for privacy-preserving data analysis”, SIGMOD, pp. 19-30, 2009.