

多次元データに対する局所差分プライバシー手法 CALM の評価

山本 悠太†

明治大学総合数理学部 先端メディアサイエンス学科 菊池研究室†

1 研究背景

デジタル化の進展に伴い、個人データの収集と分析が広範囲に行われるようになり、企業や政府機関が収集するデータの量が増大している [5]. それらのデータはターゲット広告やマイナンバーカードなどの技術に利用されている一方、同意のない情報提供などのプライバシー侵害のリスクとなるケースも多い.

データの有用性を損なわず、プライバシー侵害のリスクを軽減するための技術的対策として局所差分プライバシー [2] が注目されている. 単次元のデータに関する局所差分プライバシー手法に関しては Randomized Response[2], Count Mean Sketch[7], Optimized Unary Encoding[6], Optimized Local Hash[6] など様々な手法が提案されているが、多次元データに対する局所差分プライバシー手法に関しては十分な提案がなされていないとは言えない. その理由として、データの次元数に比例して大きなプライバシー予算がかかることが大きな課題となっていることがあげられる. 既存手法として知られる、Zhang らによる CALM[1] では、求めたい多次元データより低次元の周辺頻度を表現する、分割表の制約を無矛盾とする多次元周辺頻度をエントロピー最大となる凸最適化問題を解決することにより算出する. CALM[1] の問題として、凸最適化ソルバーの精度、ノイズの大きさや次元数によっては解不定となってしまうことが Zhang ら [1], 菊池 [3] により指摘されている.

本研究では多次元データに対する局所差分プライバシー手法とである、CALM[1] の再実装を行い、評価することを目的とする. 特に、2 次元の分割表から 5 次元の分割表の確率分布を求めるケースについて実験を行い、CALM を適用、評価した.

†Kikuchi Laboratory, Department of Frontier Media Science, School of Interdisciplinary Mathematical Science, Meiji University.

2 先行研究

2.1 局所差分プライバシー

局所差分プライバシー (Local Differential Privacy, LDP) は、データが収集される前に個人のプライバシーを保護するための枠組みである. 具体的には、データ提供者が自身のデータにランダム化処理を施すことで、データ収集者に渡される情報が個々の元データを推測することが難しくなるように設計されている.

形式的には、任意の入力 $x, x' \in X$ と任意の出力 $y \in Y$ に対して、メカニズム M が以下を満たすとき、 M は ϵ -局所差分プライバシーを満たす：

$$\Pr[M(x) = y] \leq e^\epsilon \Pr[M(x') = y], \quad (1)$$

ここで、 $\epsilon \geq 0$ はプライバシー予算 (privacy budget) と呼ばれ、プライバシーと有用性のトレードオフを制御するパラメータである. ϵ が小さいほどプライバシーは強くなるが、出力結果の有用性は低下する可能性がある.

局所差分プライバシーの基本的な仕組みでは、データ提供者がデータを収集者に送信する前に、摂動化によって情報を保護する. これにより、収集者は統計的な分析を行うために必要な集約データを得ることができる一方で、特定の個人のデータに関する情報を直接推測することが難しくなる.

2.1.1 多次元の局所差分プライバシー

一般に想定される局所差分プライバシーは 1 次元のデータに対してのものが多く、多次元に対する局所差分プライバシーには困難点が存在するためである. 考慮する属性数が増えることにより、プライバシーコストが増大する点や、また属性間の依存関係について考慮する必要があり、計算コストも増大する点などがあげられる. 代表的な手法として LoPub[8], CALM[1] がある.

2.2 CALM

CALM[1] では多次元のデータに対し、属性ごとの頻度を計算する分割表を複数作成し、これらの分割表から各属性の周辺確率を得、そこから全体の確率分布に対する

制約条件を求める。求めた制約条件が無矛盾となり、かつエントロピー最大となるように凸最適化問題を解くことによって、各ユーザのデータが秘匿された状態で全体の確率分布を求めることができる。

以下に例を示す。今、 $n = 32$ 人のユーザ v_1 から v_{32} が、"Gender", "Age", "Marital" という $k = 3$ 次元の情報を持っている。これらの情報は Gender = {Male, Female}, Age = {Adult, Elder}, Marital = {Married, Unmarried} からなり、表 1 に示す 3 次元頻度分布から、全体確率分布 $F(V)$ を収集したいケースを考える。この時、元データを直接サーバに送信してしまうと、ユーザのプライバシーが秘匿されないため、より低い $\ell < k$ 次元の周辺頻度の集計をする分割表から表 1 の頻度分布を算出することを試みる。

表 2, 3 に分割表の例を示す。表 2 では Gender, Age の周辺頻度、表 3 では Gender, Marital の周辺頻度を表す。これらの頻度は 2 属性間の関係を表している。 x_1 から x_8 を 3 次元の属性の組み合わせだとすると、表 2 からは以下のような関係を得られ、

$$\begin{cases} x_1 + x_5 = 6 \\ x_2 + x_6 = 8 \\ x_3 + x_7 = 10 \\ x_4 + x_8 = 12 \end{cases} \quad (2)$$

表 3 からは以下のような関係を得る。

$$\begin{cases} x_1 + x_3 = 4 \\ x_2 + x_4 = 6 \\ x_5 + x_7 = 12 \\ x_6 + x_8 = 14 \end{cases} \quad (3)$$

CALM ではユーザはすべてのデータを送信するのではなく、属性を選択して送信する。例えば半分のユーザは Gender, Age について RR を行い、表 2 を推定し、残りの半分のユーザは、Gender, Marital について表 3 の推定を与える、といった属性の送信方法が考えられる。

本研究では送信するデータの選択は、サーバ側がユーザ側に対し、データ送信前に送る割り当てを行うものとする。

エントロピーを最大化する凸最適化ソルバー CVXPY[9] を用いて、 x_1 から x_8 についてエントロピー最大となるように、(2) と (3) を制約条件とした凸最適化問題を解くことで全体の確率推定を行う。

表 1: 全体の頻度分布, $k = 3$

	Married		Unmarried	
	Adult	Elder	Adult	Elder
Male	1	3	5	7
Female	2	4	6	8

表 2: Gender, Age の分割表, $\ell = 2$

	Adult	Elder
Male	6	10
Female	8	12

表 3: Gender, Marital の分割表, $\ell = 2$

	Married	Unmarried
Male	4	12
Female	6	14

3 実験

3.1 実験方法

$n = 64$ 人のユーザ v_1 から v_{64} が、 a_1, a_2, a_3, a_4, a_5 という $k = 5$ 次元の情報を持つ表 4 のようなデータセットを作成する。各属性はバイナリデータであり、ユーザの持つ各属性値は $\{0, 1\} = \{0.9, 0.1\}$ の確率でランダムに決定される。求める元の確率分布を表 5 に示す。

簡単のため、本実験ではプライバシーコスト $\varepsilon \rightarrow \infty$ のケースを考える。そのため摂動を加えた後のデータは元データと全く同じ確率分布となる。

表 6 のような 2 次元の分割表 ($\ell = 2$) から、表 5 の元データの確率分布 ($k = 5$) を算出することを考える。属性数が 5、各属性がバイナリであることから、分割表の個数は ${}_5C_2 = 10$ 個である。また、表 6 から (4) のような、5 次元の確率分布を表現する変数 x_n に対する制約条件を

表 4: 作成したデータセット

	a_1	a_2	a_3	a_4	a_5
v_1	0	1	0	0	0
v_1	0	0	0	0	0
v_2	0	0	0	0	0
...	...	...	...	...	...
v_{32}	0	0	1	0	0

表 5: 元の確率分布

$\langle a_1, a_2, a_3, a_4, a_5 \rangle$	確率
$\langle 0, 0, 0, 0, 0 \rangle$	0.500
$\langle 0, 0, 0, 0, 1 \rangle$	0.047
$\langle 0, 0, 0, 1, 0 \rangle$	0.093
$\langle 0, 0, 0, 1, 1 \rangle$	0.016
$\langle 0, 0, 1, 0, 0 \rangle$	0.141
$\langle 0, 1, 0, 0, 0 \rangle$	0.047
$\langle 0, 1, 0, 0, 1 \rangle$	0.016
$\langle 0, 1, 1, 1, 0 \rangle$	0.016
$\langle 1, 0, 0, 0, 0 \rangle$	0.109
$\langle 1, 0, 0, 1, 0 \rangle$	0.016
他	0

得ることができる。得られる表は 10 個であり、1 つの表から 4 本の式を得られるため、得られる制約条件は最大 40 個である。表 6 は a_1, a_2 の周辺頻度を求める分割表であったが、他の 2 属性の周辺頻度を求める分割表に関しては、表 8 から表 16 に記載する。また、これらの表から得られる制約条件として、(8) から (16) を記載する。

エントロピーを最大化する凸最適化ソルバーを用いて、 x_0 から x_{31} についてエントロピー最大となるように、分割表から得られた式を制約条件とした凸最適化問題を解き、元データとの確率の誤差を MAE を用いて評価する。最適化問題のソルバーには、Python 用ソルバーライブラリである CVXPY[9] を使用し、得られた解について手動で計算を行い誤差を算出した。

表 6: a_1, a_2 の分割表, $\ell = 2$

	0	1
0	51	5
1	8	0

$$\begin{cases} x_0 + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 = 51 \\ x_8 + x_9 + x_{10} + x_{11} + x_{12} + x_{13} + x_{14} + x_{15} = 5 \\ x_{16} + x_{17} + x_{18} + x_{19} + x_{20} + x_{21} + x_{22} + x_{23} = 8 \\ x_{24} + x_{25} + x_{26} + x_{27} + x_{28} + x_{29} + x_{30} + x_{31} = 0 \end{cases} \quad (4)$$

3.2 結果

得られた推定結果を表 7 に示す。 $\varepsilon \rightarrow \infty$ のケースでは解不定とはならず、2 次元の分割表から 5 次元の確率分布を算出することが可能であることが確認できた。表 5

表 7: 分割表から得た確率分布

$\langle a_1, a_2, a_3, a_4, a_5 \rangle$	確率
$\langle 0, 0, 0, 0, 0 \rangle$	0.516
$\langle 0, 0, 0, 0, 1 \rangle$	0.051
$\langle 0, 0, 0, 1, 0 \rangle$	0.082
$\langle 0, 0, 0, 1, 1 \rangle$	0.008
$\langle 0, 0, 1, 0, 0 \rangle$	0.140
$\langle 0, 1, 0, 0, 0 \rangle$	0.046
$\langle 0, 1, 0, 0, 1 \rangle$	0.005
$\langle 0, 1, 1, 1, 0 \rangle$	0.003
$\langle 1, 0, 0, 0, 0 \rangle$	0.109
$\langle 1, 0, 0, 1, 0 \rangle$	0.017
他	0

と表 7 との誤差を算出した結果、 $MAE = 0.0064$ となった。最大の誤差となったのは $x_0 = \langle 0, 0, 0, 0, 0 \rangle$ の項であり、元の確率と推定結果とで 0.016 の誤差が観測された。また、 $x_{16} = \langle 1, 0, 0, 0, 0 \rangle$ の項の推定については誤差 0 で推定することができていた。

3.3 考察

実験を通して 2 次元の分割表から 5 次元の確率分布を得るような CALM が機能することが確認できた。精度の面では、プライバシーコスト $\varepsilon = \infty$ を想定しているにもかかわらず誤差が出る結果となってしまったが、これは制約条件が重複しており、十分に値を決定しきれなかったことが原因だと推測する。また、用いたソルバーの計算過程において誤差が出た可能性も検討し、用いるソルバーを変更した実験や、計算を単純化できるケースを考えた実験を行い、さらなる検証を行うことは不可欠であると考ええる。

また、本実験ではプライバシーコスト $\varepsilon = \infty$ を想定して動作検証を行ったため、元データと全く同じデータに対し CALM を適用している。そのため摂動を加えた時の動作検証はできていない。既存研究 [1][3] では摂動を加えた際にはその大きさ次第で、最適化問題が無矛盾となる解を出せず解不定となる可能性が指摘されているため、本実験で考えた 2 次元の分割表から 5 次元の確率分布を算出するケースに関してもプライバシーコストを変化させたときの動作についてさらなる調査が必要である。

また取得する分割表の大きさについても検討の余地がある。例えば $k = 5$ 次元のデータに対して、今回は $\ell = 2$

次元の分割表を取得し、これを用いて周辺確率を算出したが、 $\ell = 3$ 次元の分割表、 $\ell = 4$ 次元の分割表を取得し、そこから周辺確率を算出する方法も存在する。このように高次元データに対しての適切な分割表の取得方法に関しては自明ではないため、この点に関して更なる実験が必要である。

4 おわりに

本研究では CALM の再実装を行い、実験を行い簡易的な評価を行ったが、CALM への理解とプログラムの実装があいまいなまま実験を進めた結果、極めて単純なケースのみを想定して実験を行うこととなった。その結果、摂動の強さを変えた時の CALM の評価は行うことができなかった。今後の実験として、プライバシーコストを変化させたときの精度の調査、分割表の次元の大きさを変化させたときの誤差の変化の調査を行い、本論文をさらに補完していく予定である。

参考文献

- [1] Zhikun Zhang, Tianhao Wang, Ninghui Li, Shibo He, Jiming Chen, CALM: Consistent Adaptive Local Marginal for Marginal Release under Local Differential Privacy, CCS' 18, 2018.
- [2] J. C. Duchi, M. I. Jordan and M. J. Wainwright, "Local Privacy and Statistical Minimax Rates," 2013 IEEE 54th Annual Symposium on Foundations of Computer Science, Berkeley, CA, USA, pp. 429-438, 2013
- [3] 菊池浩明, 高次元データの局所差分プライバシーの推定精度を向上する擬似逆行列ベース手法, Computer Security Symposium 2023, pp.282-pp.287, 2023
- [4] Movielens, (<https://grouplens.org/datasets/movielens/>, 2024 年 9 月参照)
- [5] 個人情報保護委員会, (https://www.ppc.go.jp/personal_info/legal/fundamentalpolicy/, 2024 年 12 月参照)
- [6] Tianhao Wang, Jeremiah Blocki, and Ninghui Li, Somesh Jha, Locally Differentially Private Protocols for Frequency Estimation, 26th USENIX Security Symposium, 2017.
- [7] Differential Privacy Team, Apple, Learning with Privacy at Scale, 2017.
- [8] Xuebin Ren, Chia-Mu Yu, Weiren Yu, Shusen Yang, Xinyu Yang, Julie A. McCann, and Philip S. Yu, "LoPub : High-Dimensional Crowdsourced Data Publication With Local Differential Privacy," in IEEE Transactions on Information Forensics and Security, vol. 13, no. 9, pp. 2151-2166, Sept. 2018.
- [9] CVXPY, (<https://github.com/cvxpy/cvxpy/>, 2025 年 1 月参照)

表 8: a_1, a_3 の分割表, $\ell = 2$

	0	1
0	46	10
1	8	0

$$\begin{cases} x_0 + x_1 + x_2 + x_3 + x_8 + x_9 + x_{10} + x_{11} = 46 \\ x_4 + x_5 + x_6 + x_7 + x_{12} + x_{13} + x_{14} + x_{15} = 10 \\ x_{16} + x_{17} + x_{18} + x_{19} + x_{24} + x_{25} + x_{26} + x_{27} = 8 \\ x_{20} + x_{21} + x_{22} + x_{23} + x_{28} + x_{29} + x_{30} + x_{31} = 0 \end{cases} \quad (5)$$

表 12: a_2, a_4 の分割表, $\ell = 2$

	0	1
0	51	8
1	4	1

$$\begin{cases} x_0 + x_1 + x_4 + x_5 + x_{16} + x_{17} + x_{20} + x_{21} = 51 \\ x_2 + x_3 + x_6 + x_7 + x_{18} + x_{19} + x_{22} + x_{23} = 8 \\ x_8 + x_9 + x_{12} + x_{13} + x_{24} + x_{25} + x_{28} + x_{29} = 4 \\ x_{10} + x_{11} + x_{14} + x_{15} + x_{26} + x_{27} + x_{30} + x_{31} = 1 \end{cases} \quad (9)$$

表 9: a_1, a_4 の分割表, $\ell = 2$

	0	1
0	48	8
1	7	1

$$\begin{cases} x_0 + x_1 + x_4 + x_5 + x_8 + x_9 + x_{12} + x_{13} = 48 \\ x_2 + x_3 + x_6 + x_7 + x_{10} + x_{11} + x_{14} + x_{15} = 8 \\ x_{16} + x_{17} + x_{20} + x_{21} + x_{24} + x_{25} + x_{28} + x_{29} = 7 \\ x_{18} + x_{19} + x_{22} + x_{23} + x_{26} + x_{27} + x_{30} + x_{31} = 1 \end{cases} \quad (6)$$

表 13: a_2, a_5 の分割表, $\ell = 2$

	0	1
0	54	4
1	5	1

$$\begin{cases} x_0 + x_2 + x_4 + x_6 + x_{16} + x_{18} + x_{20} + x_{22} = 54 \\ x_1 + x_3 + x_5 + x_7 + x_{17} + x_{19} + x_{21} + x_{23} = 4 \\ x_8 + x_{10} + x_{12} + x_{14} + x_{24} + x_{26} + x_{28} + x_{30} = 5 \\ x_9 + x_{11} + x_{13} + x_{15} + x_{25} + x_{27} + x_{29} + x_{31} = 1 \end{cases} \quad (10)$$

表 10: a_1, a_5 の分割表, $\ell = 2$

	0	1
0	51	5
1	8	0

$$\begin{cases} x_0 + x_2 + x_4 + x_6 + x_8 + x_{10} + x_{12} + x_{14} = 51 \\ x_1 + x_3 + x_5 + x_7 + x_9 + x_{11} + x_{13} + x_{15} = 5 \\ x_{16} + x_{18} + x_{20} + x_{22} + x_{24} + x_{26} + x_{28} + x_{30} = 8 \\ x_{17} + x_{19} + x_{21} + x_{23} + x_{25} + x_{27} + x_{29} + x_{31} = 0 \end{cases} \quad (7)$$

表 14: a_3, a_4 の分割表, $\ell = 2$

	0	1
0	46	8
1	9	1

$$\begin{cases} x_0 + x_1 + x_8 + x_9 + x_{16} + x_{17} + x_{24} + x_{25} = 46 \\ x_2 + x_3 + x_{10} + x_{11} + x_{18} + x_{19} + x_{26} + x_{27} = 8 \\ x_4 + x_5 + x_{12} + x_{13} + x_{20} + x_{21} + x_{28} + x_{29} = 9 \\ x_6 + x_7 + x_{14} + x_{15} + x_{22} + x_{23} + x_{30} + x_{31} = 1 \end{cases} \quad (11)$$

表 11: a_2, a_3 の分割表, $\ell = 2$

	0	1
0	50	9
1	4	1

$$\begin{cases} x_0 + x_1 + x_2 + x_3 + x_{16} + x_{17} + x_{18} + x_{19} = 50 \\ x_4 + x_5 + x_6 + x_7 + x_{20} + x_{21} + x_{22} + x_{23} = 9 \\ x_8 + x_9 + x_{10} + x_{11} + x_{24} + x_{25} + x_{26} + x_{27} = 4 \\ x_{12} + x_{13} + x_{14} + x_{15} + x_{28} + x_{29} + x_{30} + x_{31} = 1 \end{cases} \quad (8)$$

表 15: a_3, a_5 の分割表, $\ell = 2$

	0	1
0	49	5
1	10	0

$$\begin{cases} x_0 + x_2 + x_8 + x_{10} + x_{16} + x_{18} + x_{24} + x_{26} = 49 \\ x_1 + x_3 + x_9 + x_{11} + x_{17} + x_{19} + x_{27} + x_{27} = 5 \\ x_4 + x_6 + x_{12} + x_{14} + x_{20} + x_{22} + x_{28} + x_{30} = 10 \\ x_5 + x_7 + x_{13} + x_{15} + x_{21} + x_{23} + x_{29} + x_{31} = 0 \end{cases} \quad (12)$$

表 16: a_4, a_5 の分割表, $\ell = 2$

	0	1
0	51	4
1	8	1

$$\left\{ \begin{array}{l} x_0 + x_4 + x_8 + x_{12} + x_{16} + x_{20} + x_{24} + x_{28} = 51 \\ x_1 + x_5 + x_9 + x_{13} + x_{17} + x_{21} + x_{25} + x_{29} = 4 \\ x_2 + x_6 + x_{10} + x_{14} + x_{18} + x_{22} + x_{26} + x_{30} = 8 \\ x_3 + x_7 + x_{11} + x_{15} + x_{19} + x_{23} + x_{27} + x_{31} = 1 \end{array} \right. \quad (13)$$