コミットメントの部分開示によるLDPにおけるポイズニング攻撃防止の評価

菊池 浩明[1,a)]　村上 隆夫[2,b)]

概要：ランダムな摂動化によって値のプライバシーを保護したまま統計情報を推測する局所差分プライバシーにとって，悪意ある複数の結託者による不正値を送信するポイズニング攻撃が大きな課題となっている．これに対して，Song らは Pedersen Commitment を用いて入力値ベクトルをコミットし，サーバ側で値を選択することで，不正な摂動化を防止する harmless opening を提案している．彼らは離散値を対象とした GRR と OUE へ適用していたが，それを連続値に適用するには自明ではない．そこで，本研究では，連続値を対象とした LDP 方式である Stochastic Rounding (SR) 方式に着目し，Harmless Opening を適用した VSR を提案し，その推定精度，ポイズニング耐性，通信コストなどについて評価する．

キーワード：局所差分プライバシー，ポイズニング，コミットメント

Evaluation on Partial Opening of Committed Values to prevent poisoning attack of Local Differential Privacy

Hiroaki Kikuchi[1,a)]　Takao Murakami[2,b)]

Abstract: Local Differential Privacy (LDP) enables the estimation of statistics of personal data witout violating the privacy of indivdual users. However, poisoning attacks, in which multiple colluding malicous users submit manipulated values, pose a serious threat to the reliability of LDP protocols. To address this, Song et al. proposed “Harmless Opening”, which employs Pedersen Commitments to commit input vectors and allows the server to select values, thereby preventing malicious fake perturbation. Their approach assumes to take discrete LDP mechanisms such as GRR and OUE, but extending it to continuous values is not straightforward. In this work, we focus on the Stochastic Rounding (SR) mechanism, an LDP scheme for continuous data, and propose VSR, which integrates Harmless Opening into SR. We evaluate the proposed scheme in terms of estimation accuracy, robustness against poisoning, and communication cost.

Keywords: Local differential privacy, poisoning attack, commitment

1. はじめに

個人データの収集と活用を安全に行うプライバシー保護技術 (Privacy Enhancing Technologies; PETs) の重要性が高まっている．その代表例として，Local Differential Privay (LDP) がある．ユーザがスマートフォンなどのローカルな環境で個人データをランダムに摂動化してサーバに提供することにより，入力値の秘匿性を保証しつつ，母集団の統計推定を実現する仕組みであり，Google や Apple などのプラットフォーマーに導入されている．プライバシー保護のレベルが高い反面，悪意のあるユーザが結託して虚偽の値を送信することで統計推定を歪めてしまうポイズニング攻撃が課題になっていた．

ポイズニングに対して，不正者を検出する方式が提案されている．Cao [8] らは，推定精度を低減させてポイズ

[1] 明治大学総合数理学部
School of Interdisciplinary Mathematical Sciences, Meiji University
[2] 統計数理研究所, 産業技術総合研究所, 理研 AIP
Institute of Statistical Mathematics, AIST, RIKEN
[a)] kikn@meiji.ac.jp
[b)] tmura@ism.ac.jp

表 1: 研究の位置づけ

Task	LDP	Poisoning attack	Mitigation (detect)	Mitigation (commit)
Frequency (key)	GRR[2], OUE, OLH, Rappor	[7] [9]	[8] [12]	[1] [6]
Mean (value)	SR[3], PM[4]	[10]	[10]	Our work

ニング効果を削減する方法を提案している．Wu ら [9] は，Key-value データについてのポイズニングを分析し，分析結果に基づき，ポイズニングが効果的でない条件を提案している．Li らは [10]，データをいくつかの部分集合に分割し，それぞれの推定の違いから不正者を検出する手法を提案している．

これらに対して，Song らは [1] は，不正を検出することなく，摂動化をバイパスした入力を無効にする効率の良い方式，Harmless Opening を提案した．提案方式は，ユーザが摂動化の候補値ベクトルをコミットメントし，摂動化が選択された後で，ベクトルの一部を部分的開示し，正しく摂動化プロトコルを守っていたことを証明する．真の値が漏れることなく，定められた規則に従わない不正なコミットメントのみ破棄することができる．ただし，彼らの方式は入力が離散値に限られている GRR[2] と OUE に基づいており，多くの個人データに含まれる連続量を扱うことが出来ない．

そこで，本研究は，[1] に着目し，連続値を摂動化できるように拡張して，ポイズニングに対する安全性を保証する新しい LDP 方式を検討する．ナイーブな方式は，連続値を適当にカテゴリカル値に変換して，GRR を適用することである．しかし，離散化した値の数に応じて，HO のコミットメントのコストがかかり，推定誤差への影響も自明ではない．

連続値の摂動化には，Stoacstic Rounding (SR) [3], Piecewise Mechanisum (PM) [4] が知られている．これらの方式は，入力値の符号化を確率的に行うことで，GRR よりも高い精度で統計量を推定する．それゆえ，HO によりポイズニングを防止できれば，ロバストな LDP の幅を広げることとなる．そのためには，次の様な技術的な課題を明らかにする必要がある．

- SR, PM を Harmless Opening にできるか？
- 各方式のポイズニングに対するロバスト性能は？
- 提案方式の精度やポイズニング耐性はコストは？

以上の，本研究の位置づけを，表 1 に整理する．

2. 準備

2.1 LDP

アルゴリズム $\Psi : D \to \hat{D}$ が ϵ-LDP であるのは，全ての $x_1, x_2 \in D$ について，

$$\forall T \in \hat{D},\ Pr[\Psi(x_1) = T] \leq e^{\epsilon} Pr[\Psi(x_2) = T].$$

であるとき，かつ，その時に限る．

LDP を満たす頻度オラクル (FO) として，Generaize Randomized Response (GRR) [2], Stoacstic Rounding (SR) [3], Piecewise Mechanisum (PM) [4] を考える．

2.1.1 GRR

入力を離散値 $x \in D$, $d = |D|$ とする．GRR では，確率質量関数

$$Pr[\Psi_{GRR}(x) = y] = \begin{cases} p = \frac{e^{\epsilon}}{e^{\epsilon}+d-1} & \text{if } y = x, \\ q = \frac{1}{e^{\epsilon}+d-1} & \text{if } y \neq x, \end{cases}$$

に従って，$y \in \tilde{D}$ に摂動化する．

x の頻度は，$\frac{c(x)-nq}{p-q}$ で推定できる．ここで，$c(x)$ は摂動化された値の x の頻度である．

2.1.2 SR

入力を連続値 $x \in [a, b]$ とする．$k = 2/(b-a)$ とおいて，$\tilde{x} = -1 + k(x-a)$ で，$[-1, 1]$ の連続値にする．SR メカニズムでは，次の確率質量関数

$$Pr[\Psi_{SR}(\tilde{x}) = x'] = \begin{cases} q + \frac{(q-p)(1-\tilde{x})}{2} & \text{if } x' = -1, \\ q + \frac{(q-p)(1+\tilde{x})}{2} & \text{if } x' = 1, \end{cases}$$

で $-1, 1$ の離散値に変換する．ここで，$p = \frac{e^{\epsilon}}{1+e^{\epsilon}}$ and $q = 1 - p$ とする．

SR は，$E(\frac{x'}{p-1}) = \tilde{x}$ となるので，$\left(\frac{x'}{p-1} + 1\right)/k + a$ で平均値を推定できる．

2.1.3 PM

PM も同様に，$\tilde{x} \in [-1, 1]$ の連続量を，$s = \frac{e^{\epsilon/2}-1}{e^{\epsilon/2}+1}$ と置いて，確率密度関数

$$Pr[\Psi_{PM}(\tilde{x}) = x'] = \begin{cases} \frac{e^{\epsilon/2}e^{\epsilon/2}-1}{2e^{\epsilon/2}+1} & \text{if } x' \in [\ell(\tilde{x}), r(\tilde{x})] \\ \frac{e^{\epsilon/2}-1}{2e^{\epsilon/2}+1} & \text{otherwise,} \end{cases}$$

に従って，$[-s, s]$ の値域に摂動化する．ここで，$\ell(\tilde{x}) = \frac{e^{\epsilon/2}\tilde{x}-1}{2e^{\epsilon/2}+1}$, $r(\tilde{x}) = \frac{e^{\epsilon/2}\tilde{x}+1}{2e^{\epsilon/2}+1}$, とする．値域と区間の間には，$-s \leq \ell(\tilde{x}) \leq r(\tilde{x}) \leq s$ となる関係が成立する．

2.2 ポイズニング

Cao らは，[8] において，基本的な LDP 方式の内離散値を対象にした GRR,OUE, OLH についてポイズニングのリスクを指摘し，RPA (Random Pertubed-value Attack), RIA (Random Item Attack), MGA (Maximum Gain Attack) を提案した．RPA は LDP の出力空間からランダムに出力値を選ぶ攻撃，RIA は入力空間からランダムに入力値を選んで LDP を施す攻撃，MGA は攻撃者が予め定めた入力値（target item）に対して，ポイズニング前後の頻度の推定値の差が最大となるように出力値を選ぶ攻撃である．

Wu らは，[9] にて，Key-value 形式のデータに対する LDP 方式である，PrivKV, PCKV-GRR, PCKV-UE につ

いて，Cao らと同様にして定めたポイズニング攻撃を分析している．

次の 2 種類の不正モデルが知られている．

(1) Input Poisoning Attack (IPA). 入力を意図的に操作するが，摂動化は決められたアルゴリズムに従う．

(2) Output Poisoning Attack (OPA). 摂動化アルゴリズムを無視して，出力値を任意の値に置き換える．

2.3 Hamless Opening

Song らは，[1] にて，定義域から成るベクトルの値をコミットメントした後で集計者に一要素を選んでもらい，残りの要素の内，入力値が識別不能な範囲で，摂動化アルゴリズムに従っている不要な要素を開示する Harmless Opening (HO) を提案している．コミットメントに Pedersen Commitment $C(x,\tau) = g^x h^\tau \bmod p$ を用いた Hamless Opening (HO) をアルゴリズム 1 (VGRR) で提案している [1].

Song らは，VGRR が次の性質を満たすことを証明している．

定理 2.1 [1] ℓ, ℓ_1, ℓ_2 を，$\ell = \ell_1 + (d-1)\ell_2$ で $\ell_1 > \ell_2$ であるような整数とする．ℓ, ℓ_1, ℓ_2 による VGRR は，$\ln \ell_1/\ell_2$-LDP を満たす．

VGRR では，$p = \ell_1/\ell$ の確率で真の値 x が，x 以外の $d-1$ 個の値がそれぞれ $q = \ell_2/\ell$ の確率で選ばれるので，定義により ϵ-LDP が満たされる．

定理 2.2 [1] VGRR は識別不能性，すなわち，ユーザが送信する全データ $view$ を観測する攻撃者 $\mathcal{A}^*$ が摂動化 y から真の入力 x を推定する確率が，$Pr[\mathcal{A}^*(view; y) = x] - Pr[\mathcal{A}^*(y) = x] < nelg(\kappa)$ を満たす．ここで，κ は Pedersen コミットメントの素体の大きさで決まるセキュリティパラメータである．

$d\ell_2$ 個のコミットメントを Open しても，$\mathcal{A}$ が選んだ j が真の x かどうかは区別がつかない，すなわち，Harmless (無害) である．

定理 2.3 [1] VGRR は robust，すなわち，$Pr[S_{max} = S_{in}] > 1 - nelg(\kappa)$ を満たす．ここで，S_{max} と S_{in} は，標的アイテム集合の頻度の和で定まるスコアであり，それぞれ，その最大値と IPA 攻撃者によるスコアを表す．

通常の LDP では，OPA 攻撃者によるポイズニングがスコアを最大化する $S_{max} = S_{out}$ となるが，VGRR により OPA は検査を合格できる確率が $2^{-\kappa}$ で小さくなるので，高々 IPA のスコアにしかならない．ただし，OUE においては，d 個の 2 値の値それぞれについて，VGRR を繰り返すので，各ビットが $\{0,1\}$ のどちらかであることしか証明できないため，[*1]robust でない（[1] ではより条件の弱い semi robust であることが証明されている).

[*1] 例えば，$d = 3$ の時に，$(0,1,1)$ とか $(1,1,1)$ を送る

Algorithm 1 VGRR

Require: ユーザ $\mathcal{U}$ は入力 $x \in D$ を持ち，集計者 $\mathcal{A}$ と協力して x の摂動化 y を求める．

(1) (Commit) $\mathcal{U}$ は，ℓ_1 個の x と，$D\backslash\{x\}$ の要素はそれぞれ ℓ_2 個ずつランダムな順番で含む ℓ-次元のベクトル $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_\ell)$ をコミットした $[\eta] = (\eta_1, \ldots, \eta_\ell)$ を $\mathcal{A}$ に送る．ここで，乱数 τ_i を用いて $\eta_i = C(\mu_i, \tau_i)$ とする．

(2) (select) 集計者 $\mathcal{A}$ はランダムに $j \in \{1, \ldots, \ell\}$ を選び，$\mathcal{U}$ に送る．

(3) (open) $\mathcal{U}$ は，$j_1 = j$ となり，全ての $i = 1, \ldots, \ell$ について，$\mu_{j'} = i$ となる j' が ℓ_2 個になるような $J = \{j_1, \ldots, j_{d\ell_2}\}$ を選び，コミットメント $\eta_{j_1}, \ldots, \eta_{j_\ell}$ と対応する乱数 $\tau_1, \ldots, \tau_\ell$ を $\mathcal{U}$ に開示する．

(4) (verify) $\mathcal{A}$ は，J が，$j_1 = j$ であり，$i = 1, \ldots, \ell$ について，$\eta_{j_i} = C(\mu_{j_i}, \tau_{j_i})$ であり，全ての $x' \in D$ について，$\mu_i = x'$ となる $i \in J$ が ℓ_2 個あることをを満たしていることを確認したら，$y = \mu_j$ を受理する．

3. 提案方式

3.1 概要

Song らの Hamless Opening 方式は，ZKIP なしで摂動者の不正行為を防止している反面，離散値を対象としており，大きな値域 D に分布する連続量に適用することは自明でない．そこで，VGRR を要素技術として，D を離散化した方式，および，連続量を対象とした SR, PM への適用を検討する．

3.2 脅威モデル

ユーザの集合の内，m 人のユーザが不正者と結託し，ポイズニング攻撃を実施する不正ユーザとする．残りの n 名良質なユーザがいる．不正ユーザは，LDP による推定平均値 $\hat{\mu}$ を最大化（もしくは， 最小化）すること標的とする．すなわち，操作された推定平均値と真の平均値の差で定義された利得 (Accuracy Gain) を最大化（最小化）する．

不正ユーザは OPA，IPA 攻撃を行うと想定する．利得を最大化するためには，D の最大値 b を入力し続ければ良いが，不自然に m 名のユーザが $x = b$ を入力続けると，不正者の検出も容易である．そこで，[1] と同様に，初期値 x_i を持っている m 人のユーザが結託し，標的対象としない範囲にある時は，データの提供から離脱 (halt) することを仮定する．本研究では，不正者に閾値 θ を与え，$x < \theta$ の時は離脱することとする．

3.3 GRR-HO

基本的に VGRR である．ただし，値域 D_0 を $D_1, \ldots, D_d$ の d 個の部分集合に離散化し，$x_0 \in D$ を，$x_0 \in D_i$ となる $x = i$ に置き換えて，Algorithm 1 を適用する．

$\epsilon = \ln 2$，$D = \{0,1,2\}$ の時の例を示す．$\ell_1 = 6, \ell_2 = 2$，$\ell = \ell_1 + d\ell_2 = 10$ となり，説明の為にコミットメントの位数を 17 で行うと，

Algorithm 2 SR-HO

Require: ユーザ $\mathcal{U}$ は入力 $x \in [a,b]$ を持ち，集計者 $\mathcal{A}$ と協力して x の摂動化 y を求める．

（1） $\mathcal{U}$ は，$\tilde{x} = -1 + 2(x-a)/(b-a)$ を求め，$p^* = \ell_1/\ell = q + (p-q)$, $\ell_2 = \ell - \ell_1$ を満たす適切な ℓ_1, ℓ_2, ℓ を求める．

（2） $x = 1$ if $\tilde{x} > 0$; 0 otherwise として，$D = \{-1, 1\}$ について，Algorithm 1 を適用する．$\mathcal{A}$ が y を受理したら，$(c(y)/(p-1)+1)/k + a$ で推定平均値 $\hat{\mu}$ を得る．

j									8	
μ	0	1	1	2	1	0	2	1	1	1
τ	14	8	2	12	1	8	0	5	15	7
η	2	15	1	16	6	16	4	10	12	5
J	0	1		3		5	6			9

となったとする． $\mathcal{A}$ は $j = 8$ を選び，摂動化 $y = \mu_8 = 1$ となる．$\mathcal{U}$ は，$J = \{0,1,3,5,6,9\}$ のコミットメントだけを部分開示し，$\mathcal{A}$ は 0,1,2 の全てが $\ell_2 = 2$ 個ずつ存在していたので，摂動化 y を受理する．

定理 3.1 GRR-HO は robust である．

3.4 SR-HO

入力 $x \in [a,b]$ を，SR メカニズムに沿って，$y \in \{-1,1\}$ に変換する．

$\epsilon = \ln 2$, $D = \{0,1,2,3,4\}$, $d = 5$ の時の SR の例を示す．$p^* = e^\epsilon/(e^\epsilon+1) = 2/3$ となる．D の各値における確率は，

x	$\tilde{x}$	p^*	ℓ_1	ℓ_{-1}
0	-1	1/3	4	8
1	$-1/2$	5/12	5	7
2	0	1/2	6	6
3	1/2	7/12	7	5
4	1	2/3	8	4

となり，全てを満たすためには，$\ell = 12$ となる．

定理 3.2 SR-HO は semi-robust, すなわち，$Pr[S_{max} > S_{in}] > 1 - nelg(\kappa)$ である．

ここで，VGRR プロトコルで保証されるのは，全ての $i \in \ell$ について $\eta_{j_i} = C(-1, \tau)$ となる様な J を送ってはいないこと，$x \in [a,b]$ であることだけであり，$S_{max} = S_{in}$ は保証されない．全ての要素について，ℓ_2/ℓ の確率で選ばれていることが保証される VGRR よりも完全性の意味で弱い．

また，p^* の値は，d に依存しないで ϵ だけで決まる．

3.5 PM-HO

PM では，$\tilde{x} \in [-1,1]$ の値を，$[-s,s]$ の区間の値に摂動化する．値域 $[-s,s]$ を離散値で近似すれば，GRR とほぼ同様に，Harmless Opening が可能である．ただし，GRR では，x の一要素のみが確率 p で選ばれたのに対し，PM では $[\ell(\tilde{x}), r(\tilde{x})]$ の区間の要素が p で選択されるところが異なる．

Algorithm 3 PM-HO

Require: ユーザ $\mathcal{U}$ は入力 $x \in [a,b]$ を持ち，集計者 $\mathcal{A}$ と協力して x の摂動化 y を求める．

（1） (Commit) $\mathcal{U}$ は，$\tilde{x} = -1 + 2(x-a)/(b-a)$ で $[-1,1]$ に変換する．$s = (e^{\epsilon/2}+1)/(e^{\epsilon/2}-1)$，$p = e^{\epsilon/2}/(e^{\epsilon/2}+1)$，$\ell(\tilde{x}) = \frac{e^{\epsilon/2}\tilde{x}-1}{2e^{\epsilon/2}+1}$，$r(\tilde{x}) = \frac{e^{\epsilon/2}\tilde{x}+1}{2e^{\epsilon/2}+1}$，と置き，$\ell_1/\ell = p$, $\ell_2 = \ell - s\ell_1$ を満たす ℓ_1, ℓ_2, ℓ を求める．$[\ell(\tilde{x}), r(\tilde{x})]$ の要素を ℓ_1 ずつ，それ以外の $[-s, \ell(\tilde{x})), (r(\tilde{x}), s]$ の要素を ℓ_2 個ずつ，ランダムな順番で含む ℓ-次元のベクトル $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_\ell)$ をコミットした $[\eta] = (\eta_1, \ldots, \eta_\ell)$ を $\mathcal{A}$ に送る．ここで，乱数 τ_i を用いて $\eta_i = C(\mu_i, \tau_i)$ とする．

（2） (select) 集計者 $\mathcal{A}$ はランダムに $j \in \{1, \ldots, \ell\}$ を選び，$\mathcal{U}$ に送る．

（3） (open) $\mathcal{U}$ は，$j_1 = j$ となり，全ての $i = 1, \ldots, \ell$ について，$\mu_{j'} = i$ となる j' が ℓ_2 個になるような $J = \{j_1, \ldots, j_{d\ell_2}\}$ を選び，コミットメント $\eta_{j_1}, \ldots, \eta_{j_\ell}$ と対応する乱数 $\tau_1, \ldots, \tau_\ell$ を $\mathcal{U}$ に開示する．

（4） (verify) $\mathcal{A}$ は，J が，$j_1 = j$ であり，$i = 1, \ldots, \ell$ について，$\eta_{j_i} = C(\mu_{j_i}, \tau_{j_i})$ であり，全ての $x' \in D$ について，$\mu_i = x'$ となる $i \in J$ が ℓ_2 個あることをを満たしていることを確認したら，$y = \mu_j$ を受理する．

$x \in \{0,1,2,3,4\}$ で，$\epsilon = \ln 2$ の時，$s = (e^{\epsilon/2}+1)/(e^{\epsilon/2}-1) = (\sqrt{1}+1)/(\sqrt{1}-1) = 5.8$ である．この時，

x	$\tilde{x}$	$\ell(\tilde{x})$	$r(\tilde{x})$
0	-1	-5.8	-1
1	$-1/2$	-4	0.7
2	0	-2.4	2.4
3	1/2	-0.7	4.1
4	1	1	5.8

となる．

離散化 ℓ が十分に大きいならば，攻撃者はすべての $x \in [a,b]$ について，ℓ_2/ℓ の確率で分布していることを確認できる．また，$[\ell(\tilde{x}, r(\tilde{x})]$ の要素を ℓ_1 未満にするコミットメントを作ることも可能であるが，S_{max} を下げるだけである．したがって，次を得る．

定理 3.3 PM-HO は robust である．

明らかに，SR-HO と PM-HO は，ϵ-LDP を満たしている．識別不能性についても，要素技術 VGRR の性質より，保証される．

4. 評価

4.1 評価方法

VGRR によるポイズニング防止の効果を実験データで確認する．連続量のオープンデータとして，UCI Adult データの age 属性 $(a = 17, b = 90)$ を用いる．

標的とする統計量は，age の平均値とし，精度は MSE, ポイズニング耐性についてはポイズニング後の推定値となしの推定値の差で定める利得 gain で評価する．

- 値域を d 値に量子化することによる推定誤差はどう変

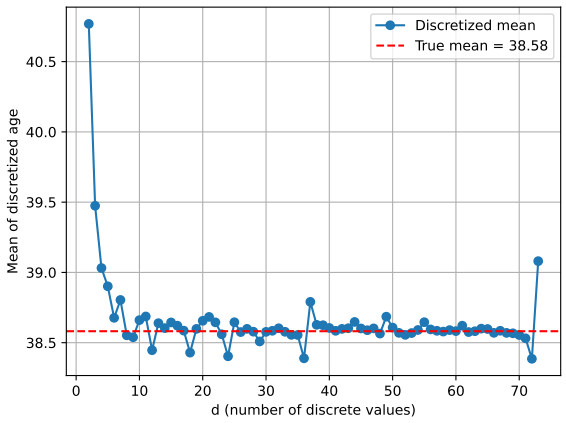


図 1: 値域離散化数 d についての平均値

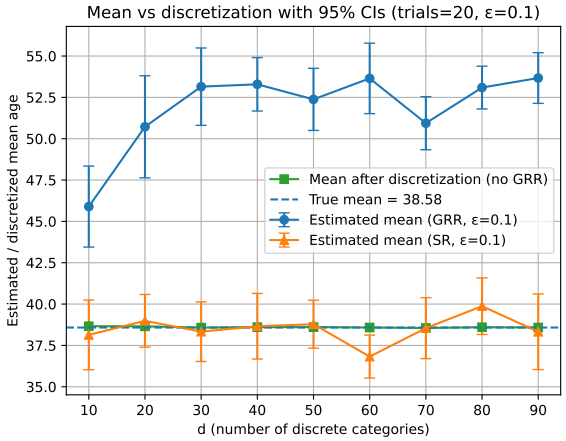


図 2: 値域離散化数 d についての GRR の推定平均値

わるか？

- Hamless Opening によるポイズニング攻撃へのロバスト性は向上するか？
- 平均値推定精度や計算コストの観点で，Harmless Opening に最適な LDP 方式は GRR, SR, PM のどれか？

4.2 評価結果

4.2.1 量子化 d

図 1 に，d 値に離散化した値域の平均値の変化を示す．$d = 10$ で真の平均値に収束しており，HO において 10 値程度でも十分な精度が期待できる．しかし，離散化した定義域について，LDP を適用したときの精度は自明ではない．図 2 に離散化数 d に対する GRR, SR の推定平均値の変化を示す．$\epsilon = 0.1$ で摂動化し，20 回試行した時の 95%の信頼区間を示している．図 1 とは逆に，d が大きくなるに連れて，真の平均 38.58 から外れている．GRR においては，MSE は $\sqrt{\frac{d}{n\epsilon}}$ で生じることが知られており，量子化により精度向上の効果と $\sqrt{d}$ に比例して生じる摂動化での誤差が相殺されたことと考えられる．

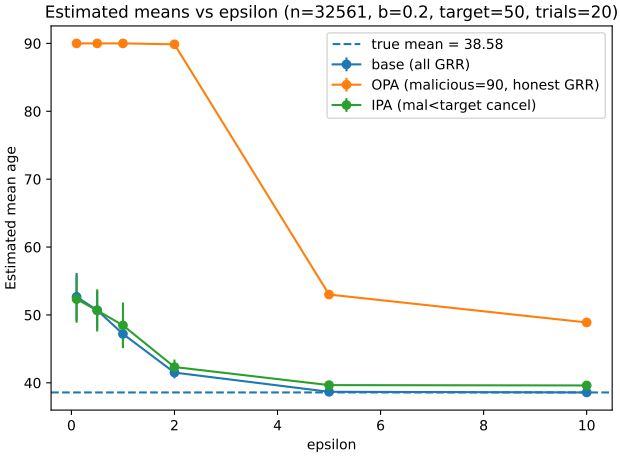


図 3: プライバシー費用 ϵ についての GRR への各種攻撃利得 ($\beta = 0.2$)

4.2.2 HO によるロバスト性

図 3 に，（生の） GRR ($d = 90 - 17 = 73, n = 32,561$) に対するポイズニング攻撃 IPA, OPA を適用した時の $\epsilon = 0.1, 0.5, \ldots, 10$ についての推定平均値の分布を示す．base はポイズニングなし，OPA では β の割合で含まれる不正者が摂動化をせずに，$y = 90$(最大値）を送信する．IPA では，β の不正者は自分の値 $x \geq \theta$ の時だけ GRR で摂動化を行い $y = \Psi_{GRR}(x)$ を送信し，θ 未満の時は参加しない．

図から，推定平均値は OPA が最も高くポイズニングに成功しているが，ϵ が 2 から減少し $\epsilon = 10$ では，48.9 まで下がり，真の値 38.58 に対して 10 増加に留まっている．一方，IPA は Base に沿って推移し，$\epsilon = 10$ の時，その差は 1.03 である．不正者が混在しても，正しく摂動化されれば影響は小さく抑えられると言える．従って，Harmless Opening により，OPA のみ禁じられれば推定は十分にロバストに行われる．

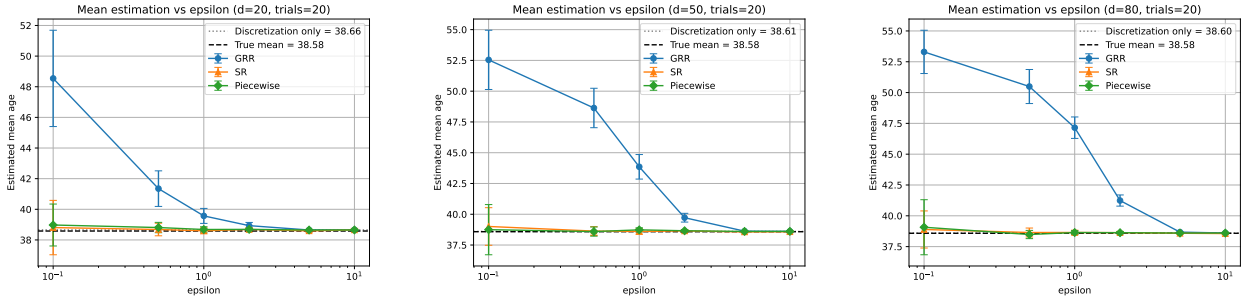
4.2.3 LDP による推定誤差

図 4 に，$d = 20, 50, 80$ の量子化における $\epsilon = 0.1, \ldots, 10$ についての GRR, SR, PM の推定平均値の分布を示す．ϵ に対して，GRR は単調に減少しているのに対し，SR, PM は安定して正確な推定を与えている．

図 5 に，$\epsilon = 0.1, 0.5, 1$ における，$d = 10, \ldots, 90$ についての GRR, SR, PM の推定平均値の分布を示す．d に対して GRR の誤差が広がっているのに対し，SR, PM は安定している．$\epsilon = 1$ に至っては，d にほぼ線形である．

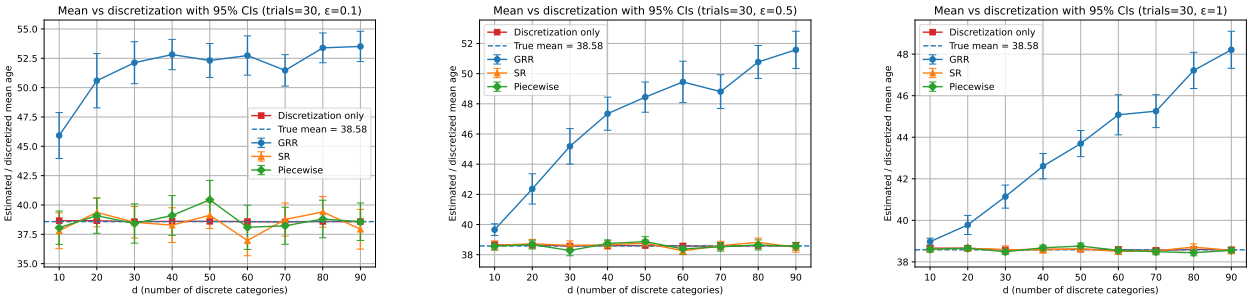
4.2.4 IPA に対する攻撃者利得

図 6 に，IPA 攻撃 ($\beta = 0.1, \theta + 40$) ににおける $\epsilon = 0.1, \ldots, 10$ についての利得の分布を示す．$d = 20, 50, 80$ の 3 種類の離散値数に対して示している．すべての d に共通して，GRR は減少し，SR と PM は増加を示し，$\epsilon = 5$ で単一の値に収束している．



(a) $d = 20$　(b) $d = 50$　(c) $d = 80$

図 4: ϵ についての推定平均値の分布



(a) $\epsilon = 0.1$　(b) $\epsilon = 0.5$　(c) $\epsilon = 1$

図 5: d 値についての推定平均値の分布

図 7 と 8 に，同様の IPA 攻撃における不正者率 $\beta = 0.1, \ldots, 0.5$, および，標的閾値 $\theta = 40, \ldots, 80$ についての利得の分布を示す．β については，線形に利得が増加する．離散化のレベルが細分化されるほど，GRR と SR, PM の差が生じている．$d = 2$ とすると，GRR と SR が同値になることと予見される．θ については，$\theta = 50$ で不規則に利得が減少している．age の分布の影響と考えられる．いずれも，θ を小さくするほど，操作した値を送信する不正者が増え，利得が高い．

4.3　計算コスト

図 9 に，Perdesen コミットメントに用いる素数 128, 256, 512 ビットについて，VGRR にかかる処理時間を示す．$\epsilon = \ln 2$, $d = 73$, $\ell = 200$ で行い，Core i9, 2.3GHz, 32GB の MacBook Pro で走る python のコードで計測している．コミットメントを python のリスト構造で実装しており，高速化の余地はあるが，より安全なサイズの素数で実現すると更に処理時間がかかる．[1] では，$d = 50, \theta = 0.99, \epsilon = 1.6$ の VGRR は，0.05 [s], 0.02 [MB] で実行できることが示されている．ただし，161 bit と 80bit で，$\kappa = 80$ bit である．

4.4　考察

Harmless Opening を適用するために，連続値を d 値に離散化する方式を検討したが，コミットメントだけでなく，推定精度と攻撃者利得にも大きく影響を及ぼすことが示された．LDP の観点では，3 つとも変わらない．ポイズニング耐性については，GRR と PM のみ robust であるが，SR については $y = \{-1, 1\}$ のどちらも ℓ_2/ℓ の確率で生じていることしか証明できず，robust ではない．一方，Harmless Opening の通信コスト，計算コストはコミットメントの数 ℓ に比例するので，SR が最小である．以上の関係を，表 2 に整理する．コスト，精度，利得は実験的に示した定性的な評価である．

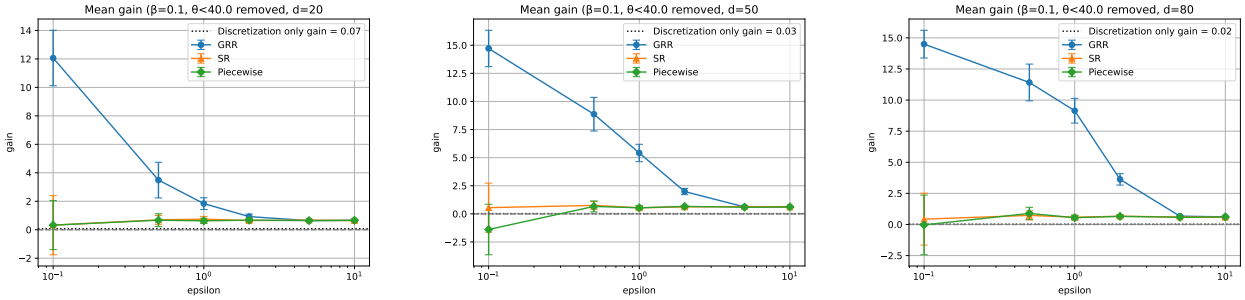
表 2: 方式の比較

性質	GRR	SR	PM
プライバシー	$\epsilon = \frac{\ell_1}{\ell_2}$-LDP	$\epsilon = \frac{\ell_1}{\ell_2}$-LDP	$\epsilon = \frac{\ell_1}{\ell_2}$-LDP
ポイズニング	robust	semi robust	robust
	$S_{max} = S_{in}$	$S_{max} > S_{in}$	$S_{max} = S_{in}$
ℓ	$\ell_1 + (d-1)\ell_2$	$\ell_1 + \ell_2$	$\ell_1 s + \ell_2 s$
コスト	high	low	high
精度	low	high	high
利得	high	low	low

なお，実験では x の定義域のみ d 値に離散化しているが，Harmless Opening を実行するためには，PM における区間 $[-s, s]$ を離散化しなくてはならない．

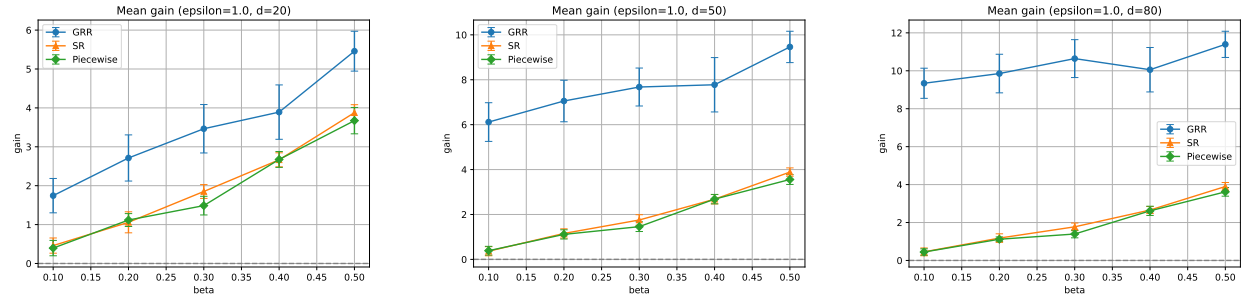
5.　おわりに

本研究では，LDP に対するポイズニング攻撃防御の観点



(a) $d = 20$ (b) $d = 50$ (c) $d = 80$

図 6: ϵ についての攻撃者利得の分布



(a) $d = 20$ (b) $d = 50$ (c) $d = 80$

図 7: 不正者率 β についての攻撃者利得の分布

から，Pedersen コミットメントを用いた Harmless Opening (HO) の枠組みを連続値処理に拡張する方式を検討した．GRR, SR, PM の 3 種類の LDP 方式に対して HO を適用し，その推定精度，攻撃耐性，および処理コストを比較評価した．

オープンデータを用いた評価の結果，GRR-HO と PM-HO は robust 性を有し，OPA に対して高い耐性を示す一方，通信コストや計算量の面で大きなオーバヘッドが生じる．SR-HO は逆に，コストが低い反面，robust 性の保証は限定的であることが明らかとなった．また，量子化レベルの選択が推定精度や攻撃者利得に大きく影響を与えることも確認された．これらの知見は，任意のシステムにおける最適な方式を設計，選択する際の指針を与えるものである．

今後の課題としては，(1) より効率的なコミットメント実装による計算コスト削減，(2) 多次元連続データへの拡張，(3) 他の攻撃モデル（例：fine-graned などの高度なポイズニング手法）への適用可能性の検証が挙げられる．

謝辞 本研究は，JST, CREST Grant Number JPMJCR21M1 と JSPS 科研費 23K11110 の助成を受けている．

参考文献

[1] S. Song, L. Xu and L. Zhu, “Efficient Defenses Against Output Poisoning Attacks on Local Differential Privacy,” in IEEE Transactions on Information Forensics and Security, vol. 18, pp. 5506-5521, 2023.

[2] P. Kairouz, K. Bonawitz, and D. Ramage, “Discrete distribution estimation under local privacy,” in Proc. 33rd Int. Conf. Mach. Learn., vol. 48, Jun. 2016, pp. 24362444.

[3] J. C. Duchi, M. I. Jordan, and M. J. Wainwright, “Minimax optimal procedures for locally private estimation,” Journal of the American Statistical Association, vol. 113, no. 521, pp. 182201, 2018.

[4] N. Wang et al., “Collecting and analyzing multidimensional data with local differential privacy,” in Proc. IEEE 35th Int. Conf. Data Eng. (ICDE), Apr. 2019, pp. 638649.

[5] H. Kikuchi, J. Akiyama, H. Gobioff, and G. Nakamura, “Stochastic voting protocol to protect voters privacy,” in Proc. IEEE Workshop Internet Appl., Jul. 1999, pp. 103111.

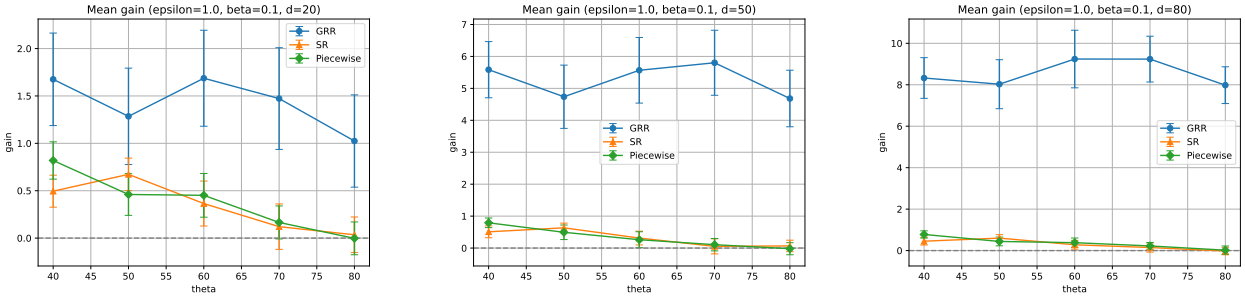
[6] F. Kato, Y. Cao, and M. Yoshikawa, “Preventing manipulation attack in local differential privacy using verifiable randomization mechanism,” in Data and Applications Security and Privacy XXXV. Cham, Switzerland: Springer, 2021, pp. 4360.

[7] A. Cheu, A. Smith, and J. Ullman, “Manipulation attacks in local differential privacy,” in Proc. IEEE Symp. Secur. Privacy (SP), May 2021, pp. 883900.

[8] X. Cao, J. Jia, and N. Z. Gong, “Data poisoning attacks to local differential privacy protocols,” in Proc. 30th USENIX Secur. Symp. (USENIX Security), 2021, pp. 947964.

[9] Y. Wu, X. Cao, J. Jia, and N. Z. Gong, “Poisoning attacks to local differential privacy protocols for Key-value data,” in Proc. 31st USENIX Secur. Symp. (USENIX Security), 2022, pp. 519536

[10] X. Li, N. Li, W. Sun, N. Z. Gong, and H. Li, “Fine-



(a) $d = 20$ (b) $d = 50$ (c) $d = 80$

図 8: 標的閾値 θ についての攻撃者利得の分布

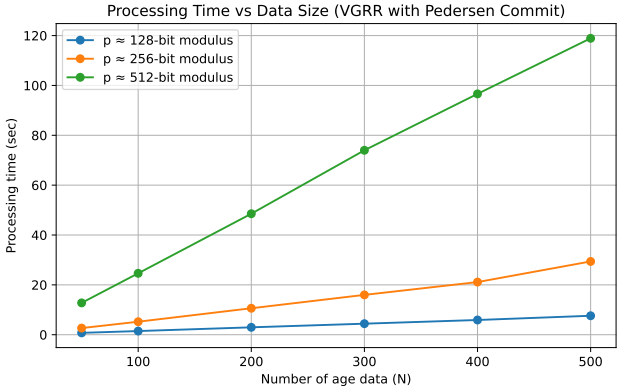


図 9: Harmless Opening GRR の処理時間

grained poisoning attack to local differential privacy protocols for mean and variance estimation," in Proc. 32nd USENIX Secur. Symp. (USENIX Security), 2023, pp. 17391756.

[11] P. Zhao et al., "An Attack-Agnostic Defense Framework Against Manipulation Attacks Under Local Differential Privacy," in 2025 IEEE Symposium on Security and Privacy (SP), pp. 3858-3876, 2025.