合成データ生成のランダム性に内在する安全性の評価

三浦 尭之[1,a)]　紀伊 真昇[1]　芝原 俊樹[1]　市川 敦謙[1]　千田 浩司[1]

概要：データ利活用の活発化に伴い，活用されるデータに含まれる個人のプライバシーを保護するプライバシー保護技術が数多く提案されている．特に近年，合成データ生成技術を用いたプライバシー保護が注目を集めている．従来の合成データ生成では，データ生成に必要な値，生成パラメータにノイズを加え，差分プライベートにすることで理論的な安全性を保証していた．しかし，企業などが所有するデータのプライバシー保護のために合成データ生成を用いる場合は，生成されたデータのみを公開し，生成パラメータは公開せずに破棄することが多い．この場合，生成パラメータを用いてデータを生成する過程にランダム性があるため，差分プライベートな合成データ生成でなくても，生成されたデータから生成パラメータおよび元データを推定することは難しい問題であると考えられる．本稿では，差分プライベートでない合成データ生成が持つランダム性がどの程度のプライバシー保護性を有しているかを考察する．理論的な評価の第一歩目として，平均と分散を生成パラメータとする正規分布に従うデータ生成が満たす安全性を差分プライバシーの考え方に基づいて評価した．さらに，評価方法を高次元データに適用する際の方向性も示した．

キーワード：合成データ生成，プライバシー保護，差分プライバシー，生成モデル

On Security of Randomness in Synthetic Data Generation

Takayuki Miura[1,a)]　Masanobu Kii[1]　Toshiki Shibahara[1]　Atsunori Ichikawa[1]　Koji Chida[1]

Abstract: With the increasing demands for the data utilization, many techniques have been proposed to protect the privacy of individuals in the data. In recent years, privacy protection techniques based on synthetic data generation has attracted much attention. Conventional synthetic data generation guarantees theoretical security by making generation parameters, which are required for the data generation, differentially private. When enterprises use synthetic data generation to protect their data, they, however, generate synthetic data by their generation parameters and discard them without disclosing them. In addition, since synthetic data generation has its own randomness, it is not easy to estimate the generation parameters and the original data from the generated data. In this paper, we theoretically discuss the difficulty of the estimation. As a first step in the theoretical evaluation, we evaluate the security of synthetic data generation by the normal distribution with the mean and the variance of the original data, referring to the concept of differential privacy. We also show the future direction of privacy-preserving data generation for high-dimensional data.

Keywords: synthetic data generation, privacy protection, differential privacy, generative model

1. はじめに

分析できるデータの量や種類の増加，深層学習をはじめとする機械学習技術の発展を背景に，データ利活用が活発化している．特に個人に関わるデータの利活用は医療分野や広告分野など様々な分野で期待されているが，これらのデータは個人のプライバシーを含むため，その取り扱いには法的・社会的な責任が伴い注意が必要である．そこで，個人のプライバシーを保護しつつ，利活用を可能にするプライバシー保護技術が盛んに研究されている [4, 13]．

過度なプライバシー保護は保護済みデータを意味のない

[1] NTT 社会情報研究所
NTT Social Informatics Laboratories
[a)] takayuki.miura.br@hco.ntt.co.jp

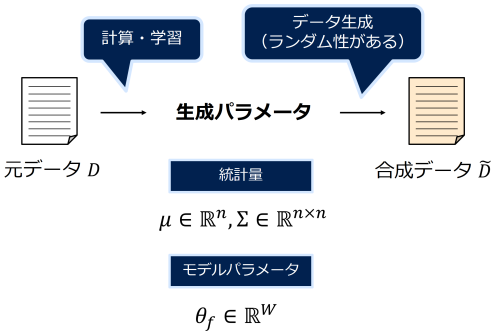

図 1 合成データ生成の概要：データセットから生成パラメータ（統計量やモデルパラメータ）を取り出し，その値に基づきランダム性をもって合成データを生成する．

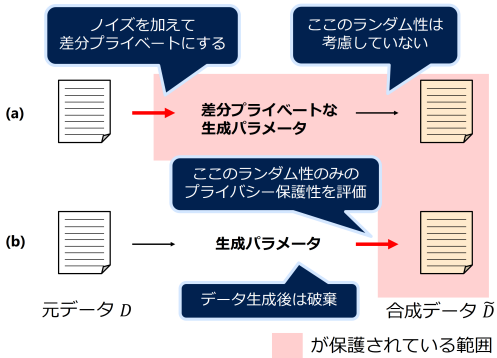

図 2 (a) 従来のプライバシー保護合成データ生成：生成パラメータの計算・学習などにランダム性を持たせて差分プライベートにすることで，そこから生成される合成データの安全性を保証．(b) 本稿で注目したランダム性：ノイズを加えるなどの保護をしない場合の合成データ生成．データ生成後は生成パラメータを破棄する．データ生成時のランダム性がそもそもどの程度のプライバシー保護性を持っているのか評価．

データにしてしまうため，k 匿名性 [13] や (ε, δ)-差分プライバシー [3,4] などの定量的な安全性指標に基づいた安全性を保証しながら，有用性と両立させることが重要である．特に，高次元のデータに対してはその両立が難しいため，高次元データに対しても有用性を保てるプライバシー保護技術として，合成データ生成技術に注目が集まっている [14,16,18]．PWS Cup 2020[*1]でも，メンバーシップ推論攻撃 [12] に対して，データセットを保護する匿名化の上位 5 チームはいずれも合成データ生成によるプライバシー保護を採用していた．

合成データ生成技術とは，保護すべきデータセットから値（本稿ではこれを**生成パラメータ**と呼ぶ）を抽出し，生成パラメータを用いて元のデータセットと同様の特徴を持つレコードやデータセットを生成する技術である（図 1）．生成パラメータとしてデータセットのもつ統計量を利用した統計量ベースの方法 [17,18] や，データセットを訓練データとし，深層学習によって得られた生成モデルを用いた方法 [5,8,11,15] などが代表的なものとしてあげられる．

これらの方式は，生成パラメータに対して，差分プライバシーを満たすようノイズを加えることによって，その安全性を保証している [1,9]．これは，統計量や生成モデルのパラメータなどの生成パラメータ自体が差分プライベートであれば，そこから生成されるデータは差分プライバシーの意味でプライバシーが保護されているという考え方である（図 2(a)）．しかし，合成データ生成はそもそも，抽出した生成パラメータからデータを生成する操作自体がランダム性を持っていて，生成されたデータから，生成パラメータや元データに関する情報を推定するのは難しい問題である．実際に，PWS Cup 2020 で合成データ生成を用いた匿名化部門上位チームの保護データは，差分プライバシーなどの保護は用いられていないにも関わらず，他のチームのメンバーシップ推論攻撃に強い耐性を見せた．

本稿では，生成パラメータの抽出時にノイズを加えるなどの保護を行わない合成データ生成がもともと持っているランダム性が，どのくらいのプライバシー保護性を有するのか評価する．すなわち，出力をレコードやテーブルとみる場合の一連の系が満たす安全性の評価を行う（図 2(b)）．

安全性の評価方法に関しては，企業などの組織が収集したデータにプライバシー保護を施して公開する場合を想定して，次のような前提があるとする．

(1) プライバシー保護をかけたいデータセット D に対して，そのデータセットに保護をかけた場合の安全性が知りたい．
(2) 作成した生成パラメータは公開せず，レコードを必要な数だけ生成した後，それは破棄される．

具体的な評価の仕方は (ε, δ)-差分プライバシーの考え方を基本としたが，本研究では，前提 (1)，(2) を反映させて，**データ固定下での (ε, δ)-差分プライバシー**（定義 3.1）を定義し，そのもとで安全性の評価を行った．定量的な評価の一歩目として，1 次元データからなるデータセットに，その平均，分散による合成データ生成を施すときのメカニズムが満たす (ε, δ) について，具体的に δ を決めたときの ε の値を理論的に導出した（定理 3.3）．

本評価は，従来のプライバシー保護と競合するものではなく，合成データ生成が自然に持っていたランダム性のプライバシー保護性を評価したものになる．今後，この観点を踏まえて，差分プライバシーによる保護などを行うことで，従来の方式で加えるノイズの量を減らすことができ，同じ安全性の保証のもと，より有用性の高いデータが生成できるようになることが期待される．

2. 準備

本節では，3 節での主定理を記述するのに必要な基本概念を紹介する．

[*1] https://www.iwsec.org/pws/2020/cup20.html

2.1 表記，定義

本稿で考えるデータセットは，大量の個人に関する情報が表形式に並べられたものである．行が各個人に関するレコードに相当し，one-hot 化などを通して，ベクトル $x_i \in \mathbb{R}^d$ として保持されていると考える．これより，N 人の個人からなるデータセットは $D = \{x_i\}_{i=1}^N \in \mathbb{R}^{N\times d}$ と表現することができる．データセットがとりうる全体の集合を $\mathcal{D}$ と表す．

プライバシー保護メカニズムとは，データセットに対するクエリの出力をランダム化したものである．ここでクエリは入力をデータセット，出力を何らかの値（統計値，レコード，機械学習のパラメータなど）とする関数 $q : \mathcal{D} \to \mathcal{Y}$ のことである．プライバシー保護メカニズムはこのクエリ q の出力に何らかの方法でランダム性を持たせた，ランダム化関数 $\mathcal{M} : \mathcal{D} \to \mathcal{Y}$ である．

2.2 合成データ生成

本稿でスコープを当てる合成データ生成は図 1 のように，次のようのプロセスで行われるものである；

(1) 元のデータセット D から生成パラメータ（統計量，モデルのパラメータ）を抽出する．
(2) それからレコードを必要な件数，ランダム性を持った手法により生成する．

特に生成パラメータが平均や分散などの統計量である統計量ベース [18]，深層学習モデルのパラメータである深層学習ベースのものがある [5, 8, 11, 15]．

2.3 多変量正規分布について

d 次元の多変量正規分布 $\mathcal{N}(\mu, \Sigma)$ とは，平均 $\mu \in \mathbb{R}^d$，分散共分散行列[*2] $\Sigma \in \mathbb{R}^{d\times d}$ をパラメータとして，次のような確率密度関数で記述される確率分布である；

$$f_{\mu,\Sigma}(x) = \frac{\exp(-\frac{1}{2}t(x-\mu)\Sigma^{-1}(x-\mu))}{\sqrt{(2\pi)^d \det\Sigma}}$$

ここで，$\det\Sigma$ は Σ の行列式を表す．

特に，1 変数の場合は μ はスカラー値，分散もスカラー値 σ^2 で表せ，

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

と書ける．

2.4 平均，分散による合成データ生成 $\mathcal{M}_G$

データセット $D = \{x_i\}_{i=1}^N \in \mathbb{R}^{N\times d}$ に対して，平均は

$$\mu_D = \frac{1}{N}\sum_{i=1}^N x_i$$

であり，分散共分散行列は j 行 k 列成分が

$$(\Sigma_D)_{jk} = \frac{1}{N}\sum_{i=1}^N ((x_i)_j - (\mu_D)_j)((x_i)_k - (\mu_D)_k)$$

の $d \times d$ 行列 Σ_D である．このとき，D を入力とし，多変量正規分布 $\mathcal{N}(\mu_D, \Sigma_D)$ に従うサンプル $x \in \mathbb{R}^d$ をランダムに一つ出力するメカニズムを $\mathcal{M}_G : \mathcal{D} \to \mathbb{R}^d$ とおく．

2.5 差分プライバシー

差分プライバシーは Dwork らによって提唱されたプライバシー保護メカニズムの安全性指標である [3, 4]．公的統計や機械学習の分野で用いられ，プライバシー保護指標のデファクトとなりつつある．

定義 2.1 (隣接データセット)**.** データセット $D \in \mathcal{D}$ に対して，それと 1 レコードだけ異なるデータセットを D の**隣接データセット**と呼ぶ．データセット D に対して，D の隣接データセット全体からなる集合を $N(D)$ とおく．隣接性の定義，つまり，“1 レコードだけ異なる”の解釈には，「データの追加削除」と「置換によるデータの書き換え」の 2 通りがあるが [7]，本稿では「置換によるデータの書き換え」を採用する．

差分プライバシーは次のように定義される．

定義 2.2 (差分プライバシー [3])**.** プライバシー保護メカニズム $\mathcal{M} : \mathcal{D} \to \mathcal{Y}$ が次を満たすとき，(ε, δ)-**差分プライバシー**を満たすという．

任意のデータセット D，任意の $D' \in N(D)$ と任意の $S \subset \mathcal{Y}$ に対して，

$$\Pr[\mathcal{M}(D) \in S] \le e^{\varepsilon}\Pr[\mathcal{M}(D') \in S] + \delta \qquad (1)$$

が成り立つ．$\delta = 0$ の場合は，ε-差分プライバシーと呼ぶ．

ε が 0 に近づけば，出力がデータセットとその隣接データセットどちらから得られたものなのか区別がつきにくくなり，ε が大きくなれば区別がつきやすくなることを意味する．ε はそういった意味で，元データセットの識別不可能性の緩和の指標になっていて，この値の小ささでプライバシー保護メカニズムの安全性を評価する．

プライバシー保護メカニズムは正しいクエリの出力に対して，適当な確率分布に従うノイズを足して実現することが主流である．ラプラス分布（両側指数分布）に従うノイズを足す場合は，$\delta = 0$ の差分プライバシーで評価できる．しかし，正規分布に従うノイズを足す場合は出現確率の低いノイズに関して，式 (1) を満たさせることができなくなる（比が発散する）ため，出現頻度が低いものは無視できる (ε, δ)-差分プライバシーで評価することが一般的である．

また，差分プライバシーを満たすメカニズムについては基本的な性質として次があげられる．

命題 2.3. (ε, δ)-差分プライバシーを満たす，メカニズム $\mathcal{M} : \mathcal{D} \to \mathcal{Y}$ を同じデータセットに対して，n 回行ったクエリ結果を並べるメカニズム

[*2] 可逆な正定値対称行列と仮定する．

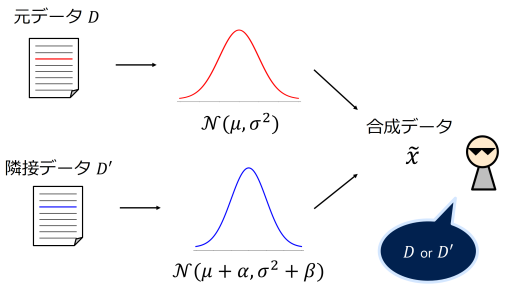

図 3 考察したい設定.

$$\mathcal{M}_n : \mathcal{D} \to \mathcal{Y}^n$$

は少なくとも $(n\varepsilon, n\delta)$-差分プライバシーを満たす.

こういった並列化されたメカニズムが満たす安全性指標は，よりタイトな評価が研究されている [6].

また，差分プライバシーは，ランダム性のあるプライバシー保護によって，出力から隣接データの区別がどれくらい行いにくくなるのかを定量的に表現した考え方である．そのため，隣接データに対して，元のクエリの出力がどれくらい変化するかということの考察は非常に重要である．そのような値をクエリの **sensitivity** と呼ぶ.

3. 合成データ生成に関する定理

本稿の目的は，合成データ生成が自然にもつランダム性がどの程度のプライバシー保護性を有するのかを調査することである．特に，合成データ生成として，2.4 項の $\mathcal{M}_G$ を用いる場合を考察する.

3.1 設定

本稿では，企業などの組織が自分らの手元にあるデータ D に合成データ生成を用いて，レコード $\tilde{x}$ もしくはテーブル $\tilde{D}$ を出力する場合を考える．そのため，

(1) プライバシー保護をかけたいデータセット D に対して，そのデータセットに保護をかけた場合の安全性が知りたい.

(2) 作成した生成パラメータは公開せず，レコードを必要な数だけ生成した後，それは破棄される.

という設定で考察を進める（cf. 図 2）．この安全性を差分プライバシーの考え方を参考に識別の難しさの観点で評価する（図 3）.

定義 3.1 (データ固定下での差分プライバシー)**. データセット D を一つ固定する**．このとき，メカニズム $\mathcal{M}$ が **データ固定下の (ε, δ)-差分プライバシー**を満たすとは，任意の $D' \in N(D)$，任意の $S \subset \mathcal{Y}$ に対して，

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S] + \delta$$

が成り立つことをいう.

注意 3.2. 通常の差分プライバシーは，「プライバシー保護メカニズム」に関する安全性であり，入力データに依存しない．しかし，本稿では前提 (1) より「保護を施すものは手元のデータ D を加工したく，またそのデータに保護をかけた場合の安全性が知りたい」という場合の考察を行うため，D 固定，つまり μ や σ^2 が固定されている（公開はされていない）と仮定する．この定義は，識別不可能性の緩和という考え方では通常の差分プライバシーと考え方は同様である．また，これらで扱うデータセット全体の集合 $\mathcal{D}$ に対して，そのデータセットの μ や σ^2 に制約を持たせれば，一般的な差分プライバシーに拡張することも可能である.

3.2 sensitivity

特に本稿の考察に必要なデータセット $D = \{x_i\}_{i=1}^N \in [l, r]^N$ の時の平均と分散の sensitivity について考察する．$x := x_N$ とおき，隣接データでは x が $y \in [l, r]$ に置き換えられたとする．平均の差は 1 レコードの差になるので，

$$\alpha := \mu' - \mu = \frac{y}{N} - \frac{x}{N}.$$

また，分散が

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2$$

と計算できることより，その差は

$$\begin{aligned}\beta &:= \sigma'^2 - \sigma^2 \\ &= \frac{y^2}{N} - \frac{x^2}{N} - \mu'^2 + \mu^2 \\ &= \alpha(y+x) - 2\alpha\mu - \alpha^2 \\ &= \alpha(2x - 2\mu + (N-1)\alpha).\end{aligned}$$

今，データの範囲は $[l, r] \subset \mathbb{R}$ としているので，レンジ幅を $R = r - l$ とおくと，

$$-\frac{R}{N} \leq \alpha \leq \frac{R}{N}.$$

また，$M = \max(r - \mu, \mu - l)$ とおくと，

$$-\frac{M^2}{N-1} \leq \beta \leq \frac{R}{N}(2M + \frac{N-1}{N}R)$$

が成り立つ．各 α, β のレンジの端点を

$$\Delta_\mu := \frac{R}{N},\ \Delta^+_{\sigma^2} := \frac{R}{N}(2M + \frac{N-1}{N}R),\ \Delta^-_{\sigma^2} := -\frac{M^2}{N-1}$$

とおいておく.

3.3 主定理

データセット D が平均 μ，分散 $\sigma^2 \neq 0$ であるとする．このとき，$0 < \delta < 1$ を一つ固定すると

$$\int_{\mu-t}^{\mu+t} f_{\mu,\sigma^2}(x)dx = 1 - \delta$$

なる $t \in \mathbb{R}$ が唯一存在するので，$x_{\pm\delta} = \mu \pm t$ とおく．本稿の主定理は次である.

定理 3.3. データセット $D = \{x_i\}_{i=1}^N \in [l,r]^N$ を平均 μ, 分散 σ^2 とする．$0 < \delta < 1$ を一つ固定し，「$\Delta_\mu < x_\delta - \mu$」と仮定する[*3].

$$L = \max(|\log \frac{\sigma^2 + \Delta^+_{\sigma^2}}{\sigma^2}|, |\log \frac{\sigma^2 + \Delta^-_{\sigma^2}}{\sigma^2}|),$$

$$m = \max_{\substack{x=x_{\pm\delta} \\ \alpha=\pm\Delta_\mu \\ \beta=\Delta^\pm_{\sigma^2}}} |\frac{(x-\mu)^2}{\sigma^2} - \frac{(x-\mu-\alpha)^2}{\sigma^2+\beta}|$$

とおく（m の式はそれぞれ端点の 2 通りずつであることに注意）．このとき，

$$\varepsilon = \frac{1}{2}(L + \max(m, \frac{\Delta_\mu(x_\delta - \mu)}{\sigma^2}))$$

とおくと，このデータセットに対するプライバシー保護メカニズム $\mathcal{M}_G$ は，データ固定下の (ε,δ)-差分プライバシーを満たす.

証明. 先に証明のスケッチをメモする；定められた δ によって, 無視してよい x のレンジ T_{out} を定め, 考察すべき出力レンジ S_{in} に焦点を絞る（Step 1）．その範囲の $x \in S_{in}$ に対して，確率密度関数の比の対数の絶対値 $g(x,\alpha,\beta)$ の上界を求めるため，式変形を行い，関数 h の上界を求める問題に帰着する（Step 2）．$h(z,\alpha,\beta)$ の各変数毎の増減を考察することで h の最大値の候補を列挙する（Step 3）．h の最大値から g の上界が求まるので，求める ε の値が決まる（Step 4）.

$D' \in N(D)$ と $S \subset \mathbb{R}$ を任意にとる.

Step 1：考えるべき出力レンジを S_{in} にしぼる

$$T_{in} := \{x \in \mathbb{R} \mid x_{-\delta} \le x \le x_{+\delta}\}$$

とおき,

$$S_{in} := S \cap T_{in}, \quad S_{out} := S \backslash T_{in}$$

とおく．このとき，$\mathcal{M}_G(D) \sim \mathcal{N}(\mu,\sigma^2)$ であるので,

$$\Pr[\mathcal{M}_G(D) \in S_{out}] \le \Pr[\mathcal{M}_G(D) \in \mathbb{R}\backslash T_{in}] = \delta$$

が成り立つ．ここで，ε に対して,

$$\Pr[\mathcal{M}_G(D) \in S_{in}] \le e^\varepsilon \Pr[\mathcal{M}_G(D') \in S_{in}] \qquad (2)$$

が成り立てば,

$$\begin{aligned}
&\Pr[\mathcal{M}_G(D) \in S] \\
&= \Pr[\mathcal{M}_G(D) \in S_{in}] + \Pr[\mathcal{M}_G(D) \in S_{out}] \\
&\le \Pr[\mathcal{M}_G(D) \in S_{in}] + \delta \\
&\le e^\varepsilon \Pr[\mathcal{M}_G(D) \in S_{in}] + \delta \\
&\le e^\varepsilon \Pr[\mathcal{M}_G(D) \in S] + \delta
\end{aligned}$$

が成り立つので，式（2）が成り立つような $\varepsilon \ge 0$ を見つければよい.

Step 2：g の上界を h の最大値に帰着する

$x \in S_{in}$ に対して，確率密度の比の絶対値の対数の関数を

$$\begin{aligned}
g(x,\alpha,\beta) &:= |\log \frac{f_{\mu,\sigma^2}(x)}{f_{\mu+\alpha,\sigma^2+\beta}(x)}| \\
&= |\frac{1}{2}\log\frac{\sigma^2+\beta}{\sigma^2} - \frac{(x-\mu)^2}{2\sigma^2} + \frac{(x-\mu-\alpha)^2}{2(\sigma^2+\beta)}|
\end{aligned}$$

とおき，その上界がもとまればそれは式（2）を満たすので，それを探す．ここで，隣接データの平均 μ'，分散 σ'^2 については．3.2 項の結果より，

$$\mu' = \mu + \alpha, \quad \sigma'^2 = \sigma^2 + \beta$$

と，$-\Delta\mu \le \alpha \le \Delta\mu$，$\Delta^-_{\sigma^2} \le \sigma^2 \le \Delta^+_{\sigma^2}$ を用いて表現できる．いま，$z := \frac{x-\mu}{\sigma}$ とおくと，

$$z_{-\delta} := \frac{x_{-\delta}-\mu}{\sigma} \le z \le \frac{x_\delta - \mu}{\sigma} =: z_\delta = -z_{-\delta}$$

であり，g は

$$\begin{aligned}
g(\sigma z - \mu,\alpha,\beta) &= |\frac{1}{2}\log\frac{\sigma^2+\beta}{\sigma^2} - \frac{1}{2}z^2 + \frac{(\sigma z-\alpha)^2}{2(\sigma^2+\beta)}| \\
&\le \frac{1}{2}|\log\frac{\sigma^2+\beta}{\sigma^2}| + \frac{1}{2}|z^2 - \frac{(\sigma z-\alpha)^2}{\sigma^2+\beta}|
\end{aligned}$$

と表せる．第一項については命題 A.1.1 より，

$$|\log\frac{\sigma^2+\beta}{\sigma^2}| \le \max(|\log\frac{\sigma^2+\Delta^+_{\sigma^2}}{\sigma^2}|, |\log\frac{\sigma^2+\Delta^-_{\sigma^2}}{\sigma^2}|) =: L$$

が成り立つので,

$$g(\sigma z-\mu,\alpha,\beta) \le \frac{1}{2}L + \frac{1}{2}|z^2 - \frac{(\sigma z-\alpha)^2}{\sigma^2+\beta}|$$

が言える．ここで,

$$\begin{aligned}
h(z,\alpha,\beta) &:= |z^2 - \frac{(\sigma z-\alpha)^2}{\sigma^2+\beta}| \\
&= |\frac{\beta}{\sigma^2+\beta}z^2 + \frac{2\alpha\sigma}{\sigma^2+\beta}z - \frac{\alpha^2}{\sigma^2+\beta}| \\
&= |\frac{\beta}{\sigma^2+\beta}(z+\frac{\alpha\sigma}{\beta})^2 - \frac{\beta}{\sigma^2+\beta}\frac{\alpha^2\sigma^2}{\beta^2} - \frac{\alpha^2}{\sigma^2+\beta}| \\
&= |\frac{\beta}{\sigma^2+\beta}(z+\frac{\alpha\sigma}{\beta})^2 - \frac{\alpha^2}{\beta}|
\end{aligned}$$

とおき，h の最大値を考えればよい.

Step 3 - 1：h の最大化（変数 z について）

z に関しては 2 次関数であるので，その絶対値が最大値をとりうるのは，頂点か両端点の 3 通りである（図 4）．今，頂点は $z = -\frac{\alpha\sigma}{\beta}$ であるが，頂点が最大値になることのそもそもの必要条件が，それが定義域に入っていることなので

$$-z_\delta \le -\frac{\alpha\sigma}{\beta} \le z_\delta$$

が条件である．よって，頂点の最大値は

[*3] この仮定は，データセットのレコード数 N が大きければ自然に満たせる条件である.

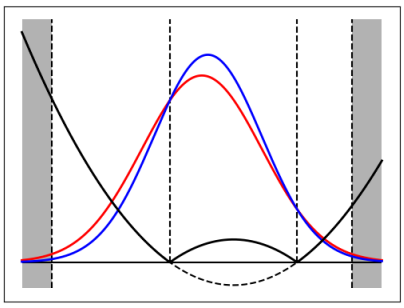

図 4　証明の中に出てくる関数 g のイメージ．
赤線：確率密度関数 f_{μ,σ^2}，**青線**：隣接データの確率密度関数 $f_{\mu+\alpha,\sigma^2+\beta}$，**黒線**：その比の対数の絶対値関数 g，この黒線関数が δ によって決まるレンジ内でとる最大値を考える．2 つの正規分布の密度関数は分散が異なれば，ちょうど二点で交わるがその点で g は 0 になる．

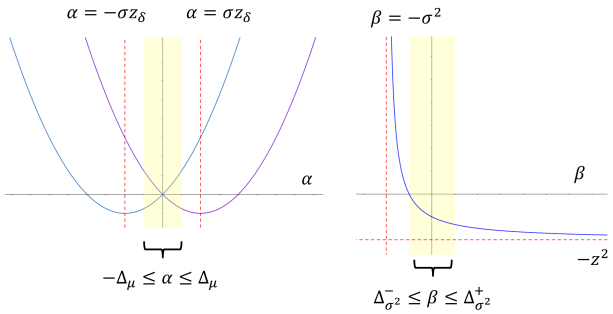

図 5　左が α の関数と見たときの関数 h の図（$z = z_\delta$ のときと，$z = -z_\delta$），右は β の関数を見たときの関数 h の図，それぞれレンジ内では単調関数になっている．

$$h(-\frac{\alpha\sigma}{\beta}, \alpha, \beta) = |\frac{\alpha^2}{\beta}| \le |\frac{\alpha z_\delta}{\sigma}| \le \frac{\Delta_\mu z_\delta}{\sigma}$$

である．以降，端点を代入した場合，つまり $z = \pm z_\delta$ とし考察する．

Step 3 - 2：h の最大化（変数 α について）

α について考えると

$$h(z, \alpha, \beta) = |\frac{1}{\sigma^2+\beta}(\alpha - \sigma z)^2 - z^2|$$

と 2 次関数になっているが，仮定 $\Delta_\mu < x_\delta - \mu$ より，頂点の軸 $\alpha = \sigma z_\delta$ は α のレンジからだいぶ離れているので，図 5 より単調であることがわかる．よって，命題 A.1.1 より，最大値は端点にあることがわかる．

Step 3 - 3：h の最大化（変数 β について）

最後に β についてみると，

$$h(z, \alpha, \beta) = |\frac{1}{\sigma^2+\beta}(\alpha - \sigma z)^2 - z^2|$$

で双曲線になっているが，いま，β の定義より，

$$-\sigma^2 < \Delta^-_{\sigma^2}$$

が成り立つので，これも図 5 より，単調性がわかり，命題 A.1.1 を適用し，その最大値は β の端点にあることがわかる．

Step 4：g の上界を踏まえて ε の値を導出

以上の考察より，求める上界の候補は，z が頂点をとる場合と，各変数 z, α, β がそれぞれ端点をとる場合であることがわかる．よって

$$\frac{1}{2}(L + h(\pm z_\delta, \pm\Delta_\mu, \Delta^\pm_{\sigma^2}) \text{ or } \frac{\Delta_\mu z_\delta}{\sigma})$$

の中の最大値が求める ε である．

□

この定理はレコード出力の場合を考察したが，テーブル出力の場合は次の系が命題 2.3 より自明に成り立つ．

系 3.4. メカニズム $\mathcal{M}_G$ を n 回使い，レコード数 n のデータセットを出力する場合を考える．このとき，メカニズムは $\delta > 0$ に対して，上記 ε を用いて $(n\varepsilon, n\delta)$-差分プライバシーを満たす．

3.4　一般次元の場合の考え方

定理 3.3 は 1 次元のデータに関するものであったが，2 次元以上の高次元の場合も同様の方式で証明できることが期待される．より複雑な問題であるため明示的な解は今後の課題とするが，問題の定式化を定理 3.3 の流れで整理する．

1 レコードが d 次元ベクトルで表現されるデータセット $D \in \mathbb{R}^{N\times d}$ を考える．隣接データセット $D' \in N(D)$ に対して，その平均を $\mu' \in \mathbb{R}^d$，分散共分散行列 $\Sigma' \in \mathbb{R}^{d\times d}$ とし，次のような関数 $g : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d\times d} \to \mathbb{R}$ を考える；

$$\begin{aligned} g(x, \mu', \Sigma') &:= |\log \frac{f_{\mu,\Sigma}(x)}{f_{\mu',\Sigma'}(x)}| \\ &= \frac{1}{2}|\log \frac{\det \Sigma'}{\det \Sigma} + {}^t(x-\mu')\Sigma'^{-1}(x-\mu') \\ &\quad -{}^t(x-\mu)\Sigma^{-1}(x-\mu)|. \end{aligned}$$

このとき，$0 < \delta < 1$ を一つ固定すると，ある $t > 0$ が唯一存在して，

$$\int_{U_\delta} f_{\mu,\Sigma}(x)dx = 1 - \delta$$

が成り立つ．ここで，$U_\delta = f^{-1}_{\mu,\Sigma}([t, \infty))$ である．このとき，

$$\varepsilon = \sup_{x\in U_\delta, \mu', \Sigma'} g(x, \mu', \Sigma')$$

とおくと，これによるメカニズム $\mathcal{M}_G$ は，データ固定下の (ε, δ)-差分プライバシーを満たす．この最適化問題が解ければ，一般次元のデータに対する合成データ生成 $\mathcal{M}_G$ が満たすデータ固定下の差分プライバシーの ε, δ の値を導出することができる．

4.　数値計算

4.1　使用したデータ，計算結果

本節では具体的な数値を使って，定理 3.3 で導出された

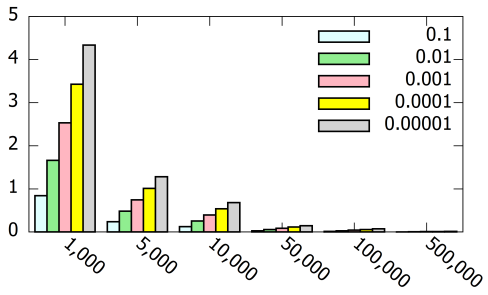

図 6 身長データに対する ε の値.

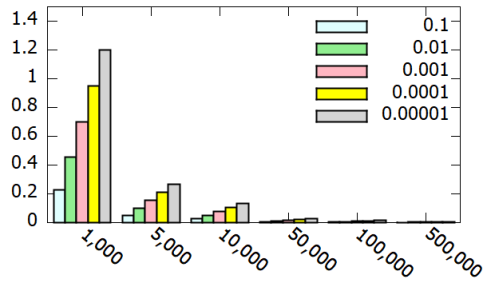

図 7 体重データに対する ε の値.

ε の値を具体的に計算する．実際のデータとして，2020 年度の学校保健統計調査における 17 歳男性の身長と体重を用いる [19]．データの概要は下記である；

- 身長：$\mu = 170.7, \sigma^2 = (5.86)^2, [l, r] = [100, 200]$
- 体重：$\mu = 62.6, \sigma^2 = (11.01)^2, [l, r] = [40, 120]$
- それぞれサンプル数 N は不明.

正確なサンプル数が不明のため，本稿では $N = 1000, 5000, 10000, 50000, 100000, 500000$ のそれぞれの場合を，また，もう一つのパラメータとして $\delta = 0.1, 0.01, 0.001, 0.0001, 0.00001$ の 5 通りで計算を行った．

結果は，身長データは図 6 で体重データは図 7 である．正確な値はそれぞれ，表 A·1 と表 A·2 に記した．それぞれのデータにおいて，ε の値は，N が大きくなれば小さくなり，δ が小さくなれば大きくなることがわかる．

4.2 考察

1 レコード当たりの出力の場合は，N が大きく，つまり元のデータセットが大きくなれば，安全性が高くなり，また，無視してよい出力の範囲を狭く（δ の値を大きく）すれば，安全性が高くなることが確かめられた．

次に，入力データの数だけ出力し，テーブル形式でデータを出力する場合を考える．この場合，系 3.4 より，ほとんどの場合 δ が 1 を超えてしまうが，これは式 1 より，意味がなくなることがわかる．$\delta > 1$ であるかどうかを無視してもすべての場合で，$\varepsilon > 100$ となり，一般的に許容されている大きさを大幅に超える結果になっている．その原因としては，用いた系 3.4 が差分プライバシーの合成則としては最もナイーブなものであり，極端な可能性の最悪ケースが出力され続けた場合を評価に入れているからであると考えられる．そのようなケースが起きる確率は低いため，あらかじめ決めた δ に対して，そのようなケースを無視するように考察する領域（3.4 項における U_δ）を定めることで，よりタイトな ε を導出することができると考えられる．

5. まとめ

本稿では合成データ生成について，生成パラメータにプライバシー保護を施す以前に，そもそものランダム性から生じるプライバシー保護性（データ固定下の差分プライバシー性）を考察した．特に，身長や体重のような 1 次元のデータの生成について理論的な ε の値の評価を与えた．この考え方は，既存のプライバシー保護と競合することはない．そのため，今後，合成データ生成を用いたプライバシー保護の安全性を評価にこの観点を組み合わせることで，有用性を相対的に上げられることが期待できる．最後に本研究の続きとして，今後考察が求められる要件をまとめる；

- α, β の関係を精密に見たバウンド
  定理 3.3 における，sensitivity の α, β の間には
  $$\beta = \alpha(2x - 2\mu + (N-1)\alpha)$$
  という関係がある．本稿ではそれぞれが独立な変数として上界を導出したが，より精密な上界はこの関係式を踏まえて計算する必要がある．
- テーブル出力
  複数のレコードを出力する場合，系 3.4 だと ε が許容できる値にならなかった．よりタイトな composition 定理による評価 [6] や，Concentrated 差分プライバシーや Rényi 差分プライバシーなどの考え方を用いた評価 [2, 10] は今後の課題である．
- データ固定下の緩和
  本稿では，データ固定下での差分プライバシーを定義し，その評価を行ったが，データセットの平均や分散のレンジに制約が設定できれば，通常の (ε, δ)-差分プライバシーに拡張することが可能である．
- 扱うデータの高次元化
  定理 3.3 では 1 次元のものに対する理論的評価を行ったが，実際のデータは高次元のベクトルである．高次元空間では最大値の探索範囲が広く本稿のように明示的に書くことは難しい．しかし，目的関数は明示的なので，最適化問題として定式化し計算機で解くことは今後の課題である．
- 深層学習モデルへの適用
  本稿の考察は合成データ生成が深層学習モデルによるものである場合も同様に行える．ただ，ニューラルネットワークによるデータの生成は，単純な正規分布に従う合成データ生成と比べると複雑なため，より詳細な考察は今後の課題である．

参考文献

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

[2] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–658. Springer, 2016.

[3] Cynthia Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pp. 1–12. Springer, 2006.

[4] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, Vol. 9, No. 3-4, pp. 211–407, 2014.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, Vol. 27, , 2014.

[6] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pp. 1376–1385. PMLR, 2015.

[7] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pp. 193–204, 2011.

[8] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[9] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the nist contest: A scalable and general approach to differentially private synthetic data. *arXiv preprint arXiv:2108.04978*, 2021.

[10] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, 2017.

[11] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pp. 1530–1538. PMLR, 2015.

[12] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2017.

[13] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 05, pp. 557–570, 2002.

[14] Shun Takagi, Tsubasa Takahashi, Yang Cao, and Masatoshi Yoshikawa. P3gm: Private high-dimensional data release via privacy preserving phased generative model. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pp. 169–180. IEEE, 2021.

[15] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *Advances in Neural Information Processing Systems*, 2019.

[16] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM Trans. Database Syst.*, Vol. 42, No. 4, October 2017.

[17] 伊原一, 田中雅行, 北林三就. 一般用ミクロデータ就業構造基本調査版の概要〜系統抽出による疑似標本データ〜. 日本人口学会第 70 大会, `https://www.nstac.go.jp/services/society_paper/30_06_03.pdf`, 2018.

[18] 岡田莉奈, 正木彰伍, 長谷川聡, 田中哲士. 統計値を用いたプライバシ保護擬似データ生成手法. コンピュータセキュリティシンポジウム 2017 論文集, 第 2017 巻, oct 2017.

[19] 文部科学省. 学校保健統計調査, 都道府県別身長・体重の平均値及び標準偏差. `https://www.e-stat.go.jp/stat-search/files?page=1&query=%E8%BA%AB%E9%95%B7%E3%83%BB%E4%BD%93%E9%87%8D%E3%81%AE%E5%B9%B3%E5%9D%87%E5%80%A4&layout=dataset&stat_infid=000032108465`, 2021.

各 URL は，2021 年 8 月 23 日に著者が確認した．

付　　録

A.1　関数の評価に関する基本的な命題

命題 A.1.1. $f : [a, b] \to \mathbb{R}$ を単調連続関数とする．このとき，任意の $x \in [a, b]$ に対して

$$|f(x)| \leq \max(|f(a)|, |f(b)|)$$

が成り立つ．

証明. 単調増加として一般性を失わない．$c \in (a, b)$ で $f(c) = 0$ となるものがある場合，f の単調性より

$$|f(x)| \leq |f(a)| \text{ if } x \leq c, \quad |f(x)| \leq |f(b)| \text{ if } x > c.$$

よって，題意の不等式が成り立つ．

また，任意の $c \in (a, b)$ で $f(c) \neq 0$ となる場合は，関数 $|f|$ も単調になるので，どちらかの端点が最大値になる． □

A.2　数値計算で得られた数値

表 6 と表 7 で示した具体的なデータを記す．

表 A·1　身長データに対する ε の値.

$N \backslash \delta$	0.1	0.01	0.001	0.0001	0.00001
1,000	0.84078	1.66091	2.53321	3.42840	4.33639
5,000	0.23742	0.48237	0.74289	1.01025	1.28143
10,000	0.12544	0.25595	0.39478	0.53724	0.68174
50,000	0.02629	0.05384	0.08315	0.11322	0.14373
100,000	0.01322	0.02710	0.04185	0.05699	0.07235
500,000	0.00266	0.00545	0.00842	0.01146	0.01455

表 A·2　体重データに対する ε の値.

$N \backslash \delta$	0.1	0.01	0.001	0.0001	0.00001
1,000	0.22506	0.45478	0.69809	0.94731	1.19978
5,000	0.04893	0.09949	0.15305	0.20791	0.26348
10,000	0.02474	0.05034	0.07746	0.10524	0.13338
50,000	0.00499	0.01017	0.01564	0.02126	0.02694
100,000	0.00250	0.00509	0.00783	0.01064	0.01349
500,000	0.00050	0.00102	0.00157	0.00213	0.00270