段階的学習を用いたプライバシ保護型深層生成モデル

高木　駿[†,††]　高橋　翼[††]　曹　洋[†]　吉川　正俊[†]

† 京都大学情報学研究科　〒 606–8501 京都府京都市左京区吉田本町 36-1
†† LINE 株式会社　〒 160–0022 東京都新宿区新宿 4-1-6 JR 新宿ミライナタワー 23 階
E-mail: †takagi.shun.45a@st.kyoto-u.ac.jp, ††tsubasa.takahashi@linecorp.com,
†††{yang,yoshikawa}@i.kyoto-u.ac.jp

あらまし　近年，大規模なデータセットを用いた機械学習とその活用が様々な領域やサービスで活用されている．しかし，それらのデータセットは個人のプライバシが問題になる場合に公開が難しく，活用が促されない．深層生成モデルは学習データと同じ特徴を持ったデータを生成するように学習するモデルであり，データそのものや匿名化済みデータの代わりに学習済み深層生成モデルを公開することが考えられる．しかし，その生成能力の高さから個人のデータを復元してしまう可能性があり，同様にプライバシ漏洩が問題になる．そこで，差分プライバシと呼ばれる厳密なプライバシ基準を取り入れた深層生成モデルを考案する．VAE や GAN といった既存深層生成モデルに既存研究である DP-SGD を単純に適応することで，差分プライバシを満たすことができる．しかし，深層生成モデルを学習する際のプライバシを保護するための雑音が原因で，生成データの質が悪くなってしまう．そこで，この論文では，段階的な学習を行うことで，その雑音の影響を軽減した新しい生成モデル Privacy Preserving Phased Generative Model (P3GM) を提案する．P3GM が差分プライバシを満たすことを示し，実験的に，同じプライバシ保護の強度の場合に既存の手法よりも質の良いデータが生成できることを示した．

キーワード　深層学習，差分プライバシ，生成モデル，プライバシ保護型データ合成

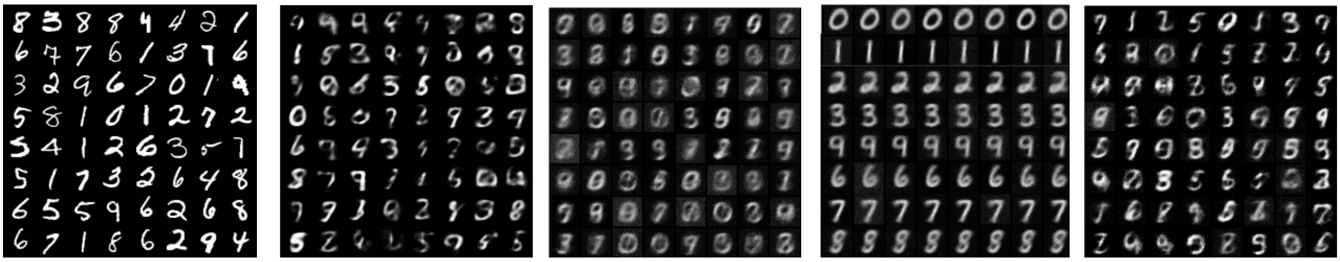
1　は じ め に

近年，電子カルテなどの情報の電子化により大規模なデータセットが増えてきている．機械学習技術の発達も伴い，そのようなデータセットを用いた機械学習による応用が期待されている．しかし，そのようなデータセットは個人のプライバシに関わる情報をしばしば含むために，公開が難しい．プライバシ保護の方法として，匿名化が考えられるが，匿名化は他の情報との組み合わせで再特定されてしまう恐れがあるためにプライバシ保護として厳密ではないことが知られている．それらの問題を受けて，近年，差分プライバシ [1] と呼ばれるプライバシ基準が広く認められてきている．もし計算機構が (ϵ,δ)-差分プライバシを満たす場合，計算機構による出力を公開したとしても，(ϵ,δ) で示される程度に個人のプライバシが厳密に保護される．直感的には，出力を見たとしても出力に加わった乱数に基づく雑音のために，データセットに任意の個人の情報が含まれていたかが推測が難しいことが保証されている．この論文では，プライバシを保護しながらデータセットを公開する方法の一つとして，差分プライバシに基づく方法を提案する．

差分プライバシを満たしながらデータセットを公開する単純な方法の一つとして，各データに雑音を加えて出力することが考えられる．しかし，一般的にデータのドメインの大きさは属性の数に従って指数的に大きくなってしまうために，差分プライバシの性質上，無視できない大きさの雑音が必要になってしまう [2]．その場合，データそのものよりも加わった雑音が大半を占めてしまうために，有用なデータを出力することができない．そこでこの論文では，データそのものに雑音を加えるのはなく，データの特徴をニューラルネットワーク（NN）を用いて学習して復元する（生成モデル）ことを考える．この場合，元のデータセットの各データに対応するもの一つ一つのデータを生成するというよりも，データセット全体として，同じ特徴を持つデータセットを復元する．データセット全体からデータを復元しようとするため，差分プライバシに必要な雑音の影響を比較的受けにくくなると考えられる．

近年，NN を用いた生成モデルが脚光を浴びており，その中でも Generative Adversarial Networks (GAN) [3] と Variational AutoEncoder (VAE) [4] がその生成能力の高さから注目されている．GAN と VAE は学習を繰り返すことで，学習データと同じような特徴を持ったデータを生成することが可能になる．一般的に NN では学習毎に確率的勾配降下法によってパラメータを更新する．つまり，NN のパラメータが個人のプライバシに関する情報を持ち得る．そこで，パラメータの更新時に雑音を加えることで，差分プライバシを満たすことができる．その場合，学習済みモデルから，学習データに含まれる個人の情報が推測されにくいことが保証される．Abadi [5] らは Differential Private Stochastic Gradient Descent (DP-SGD) と呼ばれる確率的勾配降下法に雑音を加える枠組みを提案した．この論文で提案するモデルは VAE と DP-SGD を基礎としている．

既存研究でも，同じような目的で差分プライバシを満たす生成モデルはいくつか提案されている [2] [6] [7] [8]．Zhang ら [2]



(a) MNIST(original)　(b) VAE　(c) DP-VAE　(d) DP-GM [6]　(e) 提案手法

図 1: (b) VAE [4]，(c) DP-VAE，(d) DP-GM [6]，(e) 提案手法 P3GM から生成したデータ画像．(a) は元画像である．(b)，(c)，(d) は $(1,10^{-5})$-差分プライバシを満たすモデルから生成したものである．提案手法は DP-VAE と比べてより手書き数字の特徴を捉えた画像が生成できており，DP-GM よりも多様な画像が生成できていることがわかる．

はベイジアンネットワークを訓練データを用いて構成し，そのネットワークに従って合成データを生成することを提案した．差分プライバシに必要な雑音に対して効率よく学習できるものの，計算時間の問題からベイジアンネットワークの次数に限りがあるため，複雑な依存関係を持つ高次元データの学習が根本的にできない．Xie [8] らは GAN に DP-SGD を適用することで差分プライバシを満たしている．しかし，実験では $\epsilon = 10$ が使われており，プライバシが保護されていると言える ϵ の値では実用的なデータは生成できていない．Jordon [7] らは PATE と呼ばれる枠組みを GAN に適用することで差分プライバシを満たす生成モデルを提案した．しかし，高次元データで十分な質のデータを生成できていない．GAN は多くの繊細な交互反復学習が必要となるため，差分プライバシとは相性がよくないと考えられる．なぜなら，差分プライバシでは反復によって．プライバシ漏洩が起きてしまうからである [1]．Acs [6] らは VAE に単純に DP-SGD を適用すると，データセット全体の平均のようなデータを生成するような解に陥ってしまうことを実証し，データセットをクラスタリングし，分割する枠組みを提案した．つまり，単純に言えば分割数分のデータを生成が可能になるが，分割数を増やせば各データセットの大きさが小さくなり，差分プライバシの性質上必要な雑音が増えてしまう．また，各データセットのデータの平均がそのデータセットのデータの特徴を捉えられているという保証もない．

そこで，この論文では，VAE の学習を二段階に分離することで簡略化した新しい生成モデル Privacy Preserving Phased Generative Model (P3GM) を提案する．前述したように，VAE に単純に DP-SGD を適用すると，データセットの平均データを出力する局所解に陥ってしまう．これは，VAE は埋め込みと再構築によって学習が進むと解釈されるが，その学習が複雑であることが原因にあり，差分プライバシによる雑音のために複雑な学習が進まなくなってしまうと考えた．そこで，VAE の学習の埋め込みと再構築を二段階に分離する．まず，一段階目で埋め込みを学習し，二段階目では一段階目で求めたパラメータを固定して，再構築を学習する．パラメータが固定されていることにより，VAE よりも解の探索範囲が狭まり，学習が安定すると考えられる．

一段階目の学習には主成分分析（PCA）と Expectation-Maximization(EM) アルゴリズムを用い，二段階目の学習で SGD を用いる．それぞれ差分プライバシを満たす既存技術が存在する．そして，差分プライバシの合成定理より，P3GM の学習が差分プライバシを満たすことを示すことができる．この差分プライバシの合成を厳密に行うことが重要になるが，そのために最新の技術である Rényi Differential Privacy(RDP) [9] を用いることを提案する．

この論文の目的は，より高いプライバシ保護の度合いでより質の良い人工データを生成することである．この論文では，質の良いデータとは実際のデータが持つ特徴を持っており，機械学習モデルの学習が可能であることを指す．例えば，生成したデータを用いてクラス分類機械学習モデルを学習し，そのモデルが実際のデータに対して良い精度でクラス分類ができるデータは質が良いと言う．この論文では主に二つの評価実験を行なった。一つ目は視覚的に P3GM によるデータの質を評価するために，画像データを学習し，生成を行なった．図 1 がその結果で，提案手法の生成したデータが特徴をより捉えていることがわかる．次に六つのデータセットで学習，人工データ生成を行い，五つのクラス分類モデルを生成データを用いて学習し，実際のデータでそれらのクラス分類モデルの評価を行なった．結果，同じプライバシ保護の度合いの時に，P3GM は既存研究 [2] [6] よりも人工データの質が良いことを示した．

2 準　　備

この章では，差分プライバシと P3GM の基礎となる技術について説明する．

2.1 (ϵ,δ)-差分プライバシ

差分プライバシはデータベースに含まれる個人のデータのプライバシ保護を目的としたプライバシ基準である．例えば，生成モデルをプライバシに関わる情報を含むデータセットを用いて学習をしたとする．その場合，その生成データから，個人の情報は推測されてしまうのだろうか．それは，攻撃者の事前知識や，データベースの内容に依って，できる可能性もあるし，できない可能性もあるとしか言えなく，プライバシが保護されている保証はない状態なのである．そこで，Dwork は差分プラ

イバシと呼ばれるデータベースに含まれる個人のプライバシを保護するために満たされるべき基準を提案し，今やその基準はあらゆる攻撃者やデータベースに対してプライバシの保護ができる，強力な基準であることが広く認められている．つまり，差分プライバシを満たす学習を行うことで，個人の情報が推測されにくいことが保証される．

2.1.1 定　　義

ここではその定義を述べる．そのためにまず，記法を導入する．X をレコードのドメイン，レコード $x \in X$ を個人の情報を含むデータ，レコードの集合をデータベース $D = \{x_i\}_{i=1}^n$ とする．データベース $D \in \mathcal{D}$ を引数に取り，乱数に基づく雑音を加えた応答値 $y \in Y$ を返す計算機構を M と置く．ここで取りうるデータベースの集合を $\mathcal{D}$，クエリ応答値に雑音を加えた結果得られる値の集合を Y とした．2 つの同じ大きさのデータベース D, D' において，同一でないレコードの数が一つの場合，D, D' は隣接しているという．このとき，$\epsilon \in \mathbb{R}^+$ について差分プライバシは以下のように定義される．

定義 1. *(ϵ,δ)-差分プライバシ*　隣接する任意のデータベースの組 $D, D' \in \mathcal{D}$，および任意の出力の部分集合 $S \subseteq Y$ について，

$$\frac{\Pr(M(D) \in S)}{\Pr(M(D') \in S)} \leqq \mathrm{e}^{\epsilon} + \delta \tag{1}$$

を満たす時，計算機構 M は *(ϵ,δ)-差分プライバシを満たす*という．

直感的には，M が (ϵ,δ)-差分プライバシを満たすとき，$M(D)$ を観測されたとしても，隣接するデータベースに対する出力 $M(D')$ が似ていることが保証されているため，ある一つのレコード，すなわち個人のレコードが何であるかを推測できないことを表している．

2.1.2 複数回の出力と Rényi Differential Privacy

差分プライバシは計算機構の出力が一回である場合を考えているが，機械学習タスクなどは繰り返しの出力（学習）が必要になるため，複数回計算機構を用いる場合を考える必要がある．そこで，重要な定理が合成定理である [1]．Dwork らは合成定理として，(ϵ_i, δ_i)-差分プライバシを満たす k 個の計算機構 $M_1, ..., M_k$ を用いる計算機構は $(\sum_i^k \epsilon_i, \sum_i^k \delta_i)$-差分プライバシを満たすことを証明した．しかし，これは厳密でなく，より強いプライバシ保護の (ϵ,δ)-差分プライバシを満たすことが知られている．ただし，その厳密解を求めることは#P 困難であることが示されている [10]．そこで重要なのが，より厳密な近似解を求めることである．最新の研究の動向では，各計算機構の出力でどれだけプライバシが漏洩したかをプライバシ損失変数として差分プライバシよりも厳密に測り，最後にそれらを合計し，最終的な (ϵ, δ) を求めるものである．この論文では，Rényi Differential Privacy (RDP) [9] がより厳密なことが知られているため，RDP を用いる．

定義 2 (RDP)**.** 任意の隣接データベース D, D' に対して以下を満たすとき，計算機構 M は *(α, ϵ)-RDP を満たす*という．

$$\frac{1}{\alpha - 1} \log \mathbf{E}_{z \sim M(D')} \left(\frac{\Pr(M(D) = z)}{\Pr(M(D') = z)} \right)^{\alpha} \leqq \epsilon \tag{2}$$

ただし，$\alpha > 1$ である．

さらに RDP では，以下の合成定理が成り立つ．

定理 1. 計算機構 M_1, M_2 がそれぞれ *(α, ϵ_1)-RDP*，*(α, ϵ_2)-RDP* を満たすなら，M_1 と M_2 を使う計算機構は *(α,$\epsilon_1 + \epsilon_2$)-RDP* を満たす．

差分プライバシとの関係として以下の定理が成り立つ．

定理 2. 計算機構 M が *(α, ϵ)-RDP* を満たすなら，任意の $\alpha > 1$，$0 < \delta < 1$ に対して，M は *($\epsilon + \frac{\log 1/\delta}{\alpha - 1}$,$\delta$)-差分プライバシを満たす*．

RDP はプライバシ損失をより厳密に計測しており，さらに，(ϵ, δ)-差分プライバシに変換可能である．つまり，*RDP* を用いてプライバシ損失の合成を行い， *RDP* から差分プライバシへ変換することで，より厳密な合成の計算をできる．

2.2 Variational Autoencoder

VAE では *NN* を用いた潜在変数 z の事前分布を仮定する確率モデル p_θ のパラメータ θ の最尤推定を行う．一般的には解析的に微分可能にするために，潜在変数 z の事前分布 $p_\theta(z)$ が多変量標準正規分布 $\mathcal{N}(0, I)$ に，$p_\theta(x|z)$ は多変量正規分布もしくは多変量ベルヌーイ分布に従うと仮定する．ここで尤度 $p_\theta(x)$ は *KL* ダイバージェンスを用いて以下のように変形できる．

$$\log p_\theta(x) \geqq -D_{KL}(q_\phi(z|x)||p_\theta(z)) + \int q_\phi(z|x) \log p_\theta(x|z) dz \tag{3}$$

ここで，$q_\phi(z|x)$ は，$p_\theta(z|x)$ を解析的に考えないようにするために，$p_\theta(z|x)$ を近似する分布である．一般的に右辺を変分下限と呼び L と書く．確率的勾配降下法によりこの対数尤度を最大化するように *NN* を学習する．その構造の一つの例として，$q_\phi(z|x)$ は多変量正規分布の平均と分散共分散行列を出力する *NN* を用いて，$p_\theta(x|z)$ は多変量正規分布，もしくは多変量ベルヌーイ分布の平均（分散は定数であると仮定する）を出力する *NN* を用いる構造が挙げられる．その場合，*(3)* 式の第二項の最大化は，モンテカルロ近似により，*NN* の出力とデータ x の二乗誤差の最小化と等しい．直感的には，第二項はデータの再構築をするための損失関数であり，第一項はデータ x の潜在変数 z への埋め込みを学習するための損失関数と言える．つまり，*VAE* は埋め込みと再構築を同時に学習していると解釈できる．

2.3 Differential Private EM Algorithm

DIfferential private EM (DP-EM) [11] は差分プライバシを満たす *expectation-maximization(EM)* アルゴリズムである．*EM* アルゴリズムは，尤度を最大化するパラメータを求めるアルゴリズムである．*DP-EM* は指数族で完全データの対数尤度関数の最適化が可能なモデルに適用可能であり，この論文では混合ガウス分布に適用する．混合ガウス分布は $p(x|\pi, \mu, \Sigma) = \Sigma_{k=1}^K \pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k)$ と表され，$\Sigma_{k=1}^K = 1$ を満たす．ここで K は混合数（コンポーネント数と呼ぶ）である．*DP-EM* を用いることで，差分プライバシによってプライバシ保

護をしながら，混合ガウス分布のパラメータの最尤推定を行うことができる．*EM* アルゴリズムでは *E* ステップと *M* ステップと呼ばれる操作を繰り返すことで推定値を更新する．*M* ステップでパラメータを更新するため，プライバシ漏洩が起きるのは *M* ステップであり，次のように *M* ステップで求めた最尤推定値に雑音を加えることで差分プライバシを満たすことができる．$\tilde{\pi} = \pi + (Y_1, ..., Y_K), \tilde{\Sigma}_k = \Sigma_k + Z, \tilde{\mu}_k = \mu_k + (W_1, ..., W_d)$．ここで Y, Z, W は正規分布に従う雑音である．詳しくは *[11]* を参照されたい．

2.4 Differential Private PCA

この論文では高次元データを扱うために，*principal component analysis(PCA)* による次元圧縮を行うが，*PCA* では求めた主成分がプライバシ漏洩を起こしうる．*Differential Private PCA(DP-PCA) [12]* では次のように，分散共分散行列 A に雑音を加えることで差分プライバシを満たす．$\hat{A} = A + W \quad (W \sim W_d(d+1, C_w))$．ここで W_d はウィシャート分布で d がデータの次元数，C_w は d 個の同じ固有値 $\frac{3}{2n\epsilon}$ を持つ行列である．$\hat{A}$ を求める操作は ϵ-差分プライバシを満たすため，$\hat{A}$ を用いた *PCA* も差分プライバシの *post-processing* 定理 *[1]* より，$(\epsilon, 0)$-差分プライバシを満たす．

2.5 Differential Private Stochastic Gradient Decent

P3GM の学習では *NN* を確率的勾配降下法を用いてパラメータを更新する．*Differential private stochastic gradient descent* (*DP-SGD*) *[5]* は差分プライバシを満たす確率的勾配降下法である．パラメータの更新に *DP-SGD* を用いることで *P3GM* も差分プライバシを満たすことができる．*DP-SGD* では *DP-EM* と同様にパラメータの更新が複数回行われる．*Abadi* ら *[5]* は *Moment Accountant(MA)* と呼ばれる概念を導入して差分プライバシの合成を行い，最終的な (ϵ, δ) の値を計算している．この論文では *RDP* を用いる．詳しくは *3.4* 節で述べる．

3 提案手法: P3GM

ここでは新しい生成モデルである，*Privacy Preserving Phased Generative Model*（*P3GM*）を提案する（差分プライバシを導入しない場合，*Phased Generative Model(PGM)* と呼ぶ)．単純に *VAE* に *DP-SGD* を適用してもその学習の複雑さが原因で良い解への収束が難しい．視覚的には図 *1(b)* がそれを示している．それを解決するため，*VAE* の埋め込みと再構築の学習を *P3GM* では二つの分離した段階で学習を進める．

（*1*）（埋め込みの学習）*DP-PCA* による次元圧縮と *DP-EM* による混合正規分布のパラメータの最尤推定

（*2*）（再構築の学習）埋め込みで得たパラメータを固定し，*DP-SGD* による *NN* の訓練

段階一は潜在変数の事前分布を混合正規分布を仮定して，最尤推定によって求めている．なお，高次元では最尤推定がうまく機能しないため，次元圧縮を行なっている．直感的には，データの埋め込み先の分布（つまり，潜在変数の分布）を求めている．段階二では第一段階で求めた潜在変数の分布から実際のデータを構築することを学習する．直感的にはデータの埋め込み先から実際のデータを構築することを学習している．データを生成する際は，埋め込みの学習で求めた混合正規分布を用いて潜在変数の値をサンプルし，*NN* に入力し，再構築することで人工データを生成することができる．

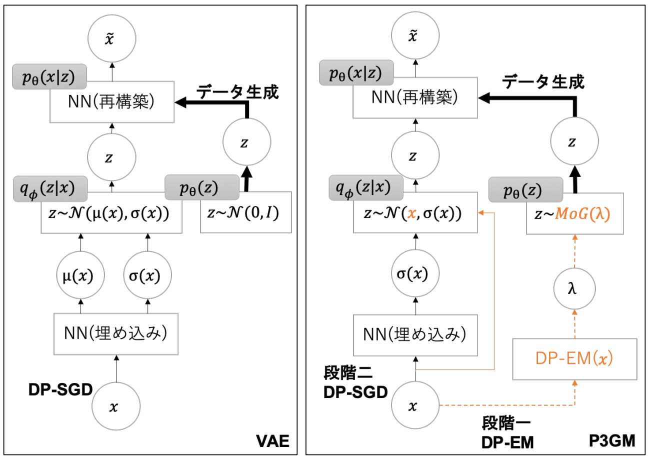


図 2: VAE（左）と P3GM（右）の流れ．$\tilde{x}$ は生成データである．点線が段階一，実線が段階二，太線がデータ生成の流れを表している．オレンジ色は VAE と異なる箇所を表す．

3.1 構　造

2.2 節で述べたように，*VAE* は再構築と埋め込みを同時に学習していると考えられる．これはもし，埋め込みがうまく学習されないと，再構築もうまく学習できないことを意味しており，逆も然りである．我々は *VAE* が *DP-SGD* でうまく学習できないのは，雑音によりこの二つの同時学習が進まないからと考えた．そこで，埋め込みと再構築の学習を二段階に分離して行うモデルを提案する．一段階目で埋め込みを学習し，埋め込みを固定して再構築を学習することにより，雑音が含まれる *DP-SGD* においても安定した学習ができると考えられる．図 *2* は *P3GM* と *VAE* の流れである．*P3GM* では段階二において，埋め込み $\mu(x)$ が省略できているのがわかる．これは段階一の学習によるものであり，段階二の *DP-SGD* による学習を安定させる．

3.1.1 段階一:DP-EM による埋め込み学習

ここでは段階一の埋め込みの学習について述べる．*(3)* 式における第一項（埋め込みに関する損失）は，データ x が与えられた時の潜在変数 z の事後分布 $p_\theta(z|x)$ を近似する $q_\phi(z|x)$ を学習するための項である．それを段階一で学習をする．直感的には，似たような特徴を持つ二つのデータは似たような二つの潜在変数の値から構築されると考えるのが自然なため，逆に似たような二つのデータは似たような二つの潜在変数の値に埋め込まれるべきである．この考えから，潜在変数 z とデータ x のドメインが同じであると仮定し，以下のようにデータ x_i が与えられた時の z の事後分布は，x_i を平均とする正規分布と仮定することは自然である．このことについては，*3.2*，*3.3* 節でさらに詳しく述べる．

$$q_\phi(z|x = x_i) = \mathcal{N}(\mu = x_i, \sigma = \sigma(x_i)) \tag{4}$$

データが高次元の場合，次元数 d から d' への次元圧縮 $f:\mathbb{R}^d \to \mathbb{R}^{d'}$ 後のデータ $f(x)$ を正規分布の平均と仮定する．次元圧縮がデータの特徴を保持しているならば，同様に $f(x)$ はデータの特徴をよく表しているため，この仮定は自然である．$\sigma(x)$ は別で求める必要があるため，段階二で学習を行う．次に，潜在変数 z の事前分布 $p_\theta(z)$ を考える．今，上で述べた仮定は，データは同じ値の潜在変数から生成されるというものであった．故に，z の事前分布と x の事前分布は同様の分布であると仮定できる．よって，$p_\theta(z)$ を $p(x)$ とする．ただし，正確な x の事前分布は求めることはできないため，混合正規分布を仮定し，*DP-EM* によってデータ x について最尤推定を行うことで混合正規分布のパラメータ λ を推定する．そして，その混合正規分布 $MoG(\lambda)$ を z の事前分布として用いる．

3.1.2 段階二:DP-SGD による再構築学習

段階二では，段階一で求めた x の埋め込みの分布 $q_\phi(z|x = x_i) = \mathcal{N}(\mu = x_i, \sigma = \sigma(x_i))$ の平均 $\mu = x_i$ と z の事前分布 $p_\theta(z)$ を固定して，再構築を学習する．*(3)* 式は入力を x_i とする場合，サンプル回数 L のモンテカルロ近似を用いて以下のように書ける．

$$-D_{KL}(q_\phi(z|x = x_i)||p_\theta(z)) + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(x = x_i|z = z^{(l)}) \tag{5}$$

ここで $z^{(l)} \sim q_\theta(z|x = x_i)$ である．第二項の最大化は，$p_\theta(x|z)$ を多変量正規分布で仮定し，*NN* でその平均値を出力すれば，x_i と出力の二乗誤差の最小化と等しくなる．第一項は第一段階で求めた埋め込み $q_\theta(z|x)$ と $p_\theta(z)$ を用いることができる．つまり，混合正規分布間の *KL* ダイバージェンスを計算すれば良い．しかし，それの解析的な計算は難しく，モンテカルロ近似では分散が大きくなってしまうため，良い近似ではなくなってしまう．そこで変分近似を用いる *[13]*．二つの混合正規分布 $f(x; \pi_a, \mu_a, \sigma_a)$ と $g(x; \pi_b, \mu_b, \sigma_b)$ 間の *KL* ダイバージェンスは以下のように近似することができる．

$$\sum_a \pi_a \log\left(\frac{\Sigma_{a'}\pi_{a'} \exp\left(-D_{KL}(\mathcal{N}(\mu_{a'}, \sigma_{a'})||\mathcal{N}(\mu_a, \sigma_a))\right)}{\Sigma_b \pi_b \exp\left(-D_{KL}(\mathcal{N}(\mu_a, \sigma_a)||\mathcal{N}(\mu_b, \sigma_b))\right)}\right) \tag{6}$$

(3) 式の第一項と第二項が微分可能なため，*DP-SGD* を用いてパラメータの更新（再構築の学習）が可能である．

3.2 PGM の解釈

我々は，*DP-SGD*（つまり，プライバシ保護のために雑音を必要とする条件下での学習）では事前学習で求めた一部のパラメータを固定し，解の探索範囲を狭めることで，精度の向上が得られると仮定している．一部のパラメータを固定して解の探索範囲を狭めることで二つの効果が得られると考える．一つ目は，少ない学習回数で収束できるということであり，二つ目は，ある程度の質の局所解を見つけることができるということである．ここでは，まずその二つの効果が *DP-SGD* に対して効果的であると考える理由を述べて，実際に *PGM* の解の探索範囲が *VAE* よりも狭くなっていることを述べる．

まず一つ目の理由について説明する．差分プライバシの合成定理より，学習回数が増えれば増えるほどプライバシ漏洩が起きてしまう．これは，最終的に一定のプライバシ保護をするためには，一回の学習当たりに多くの雑音を要することを意味する．逆に学習回数が少なければ，プライバシ漏洩の機会を減らせるため，学習当たりの雑音を減らすことができる．次に二つ目の理由について説明する．一部のパラメータを固定しているため，雑音により悪い方向に変化することがなく，そのパラメータを持つ場合の解に収束しやすくなっていると考える．パラメータを固定せずに探索をする場合，そのパラメータ自体が雑音により変化し，収束が安定しない可能性がある．

実際に *PGM* の解の探索範囲が，パラメータを固定することで *VAE* の探索範囲より狭くなることについて述べる．*VAE* では *(3)* 式において，$q_\phi(z|x)$ が多変量正規分布であると仮定し，以下を目的関数として，$\mu(x)$ と $\sigma(x)$ と θ の最適化をする．

$$\begin{aligned} L_{VAE} = &- D_{KL}(\mathcal{N}(\mu(x), \sigma(x))||p_\theta(z)) \\ &+ \int \mathcal{N}(z; \mu(x), \sigma(x)) \log p_\theta(x|z)dz \end{aligned} \tag{7}$$

PGM では同じ式において，$q_\phi(z|x)$ が平均が定数 c_x の正規分布であると仮定し，$\sigma(x)$ と θ の最適化をする．

$$\begin{aligned} L_{PGM} = &- D_{KL}(\mathcal{N}(c_x, \sigma(x))||p_\theta(z)) \\ &+ \int \mathcal{N}(z; c_x, \sigma(x)) \log p_\theta(x|z)dz \end{aligned} \tag{8}$$

この二つの損失関数を比べる．$\mu(x), \sigma(x), \theta$ があらゆる値を取り得ると仮定すると，*VAE* で得られる解は $\mu(x) = c_x$ とすることで，明らかに *PGM* で得られる解を含む．このことから，*VAE* の解を探索する範囲は *PGM* よりも広いと言える．逆に言えば，*PGM* は解の探索範囲を，*VAE* では探索していた $\mu(x)$ を定数として固定することで，狭めていると言える．

さらに，以下のように $\sigma(x)$ も定数 s_x であると仮定することで，さらに探索範囲を狭めることができると考えられる．

$$\begin{aligned} L_{AE} = &- D_{KL}(\mathcal{N}(c_x, s_x)||p_\theta(z)) \\ &+ \int \mathcal{N}(z; c_x, s_x) \log p_\theta(x|z)dz \end{aligned} \tag{9}$$

この場合，最適化する必要があるのは θ のみとなる．$s_x = 0$ と仮定すると，第一項は定数となり，オートエンコーダ *(AE)* と呼ばれるモデルになる．これらの三つの損失関数について，収束の速さと実際に探索範囲の狭い順番に速いことを第四章で実験的に示す．

3.3 議　　論

前の節で述べたように，*PGM* は *VAE* よりも探索範囲が狭いと言える．ここで議論する必要があるのが，狭めたことにより，解の質が落ちる可能性があるということである．ここでは，そのことについて述べる．今，*(3)* 式を変分下限を用いて変形すると以下のようになる．

$$\ln(p_\theta(x)) - L = D_{KL}(q_\phi(z|x)||p_\theta(z|x)) \tag{10}$$

このことから，$q_\phi(z|x)$ と $p_\theta(z|x)$ の *KL* ダイバージェンスが小さい時，対数尤度と変分下限の差が小さいことがわかる．特に $q_\phi(z|x) = p_\theta(z|x)$ の場合，その差は 0 となるため，対数尤度と変分下限は等しいことがわかる．逆に *KL* ダイバージェンスが大きいとき，変分下限は正しく対数尤度を表せていないことがわかる．つまり，変分下限の最大化において，$q_\phi(z|x)$ が正しく $p_\theta(z|x)$ を近似できている場合，対数尤度も最大化されていると考えられる．逆に近似が正しくない場合，変分下限の最大化をしても，対数尤度は大きくなっていない可能性がある．

PGM では $q_\phi(z|x=x_i)$ を平均が x_i である正規分布であると仮定している．我々は潜在変数の事前分布 $p_\theta(z)$ がデータの事前分布 $p(x)$ に等しい場合に，この仮定は妥当であると考る．これはつまり，データの生成モデルが，z がまず生成され，z の周辺のデータ x が生成されるというモデルであることを意味する．このようなモデルを考えた場合，x_i が与えられた時の z の事後分布 $p_\theta(z|x=x_i)$ は平均が x_i であると仮定できる．その場合，平均が等しい分布は似ていると考えられるため，その分布間の *KL* ダイバージェンスはある程度小さく抑えられるはずである．そして，$q_\phi(z|x)$ の分散を *NN* を用いて求めることで，よりその *KL* ダイバージェンス *((10)* 式*)* が小さくできると考えられる．これが，$q_\phi(z|x=x_i)$ の平均を x_i としても，ある程度の解の質を保証できると考える理由である．*AE(*式 *(9))* では，分散も定数であると考えるが，その場合，*(10)* 式における *KL* ダイバージェンスが一定以下にならないため，*PGM* の方がより良い解を見つけられる可能性があると考えられる．ただし，*AE* はその探索の範囲の狭さから収束はより速いと考えらえる．五章で実験的に，*PGM* の方が収束は遅いものの *AE* より良い解を見つけられていることを示す．

もう一つ議論する必要があることが，$p(x)$ を混合正規分布で近似することである．カテゴリデータや外れ値がある場合など，$p(x)$ が混合正規分布で表せない場合，この近似は荒くなってしまい，その場合，変分下限（*(3)* 式）の第一項が小さくなってしまう．なぜなら，$q_\phi(z|x=x_i)$ が x_i を平均とする分布であるためである（*VAE* では平均を固定しないため問題が生じない）．*PGM* では再構築学習時に $q_\phi(z|x=x_i)$ の分散を学習するため，第一項をある程度大きくすることができる．つまり，再構築学習によって $p(x)$ の近似の荒い部分はある程度は補正されるため，問題にならないと考えられる．

3.4 プライバシ分析

P3GM は三つの構成要素 *PCA*，*EM*，*SGD* で構成されており，各構成要素は既存技術を用いて差分プライバシを満たすことができる．それぞれ $(\epsilon_p,\ 0)$，$(\epsilon_e,\ \delta_e)$，$(\epsilon_s,\ \delta_s)$-差分プライバシを満たす時，*2* 章で述べたように，*Dwork* らの差分プライバシの合成定理 *[1]* を適用すると，*P3GM* は $(\epsilon = \epsilon_p + \epsilon_e + \epsilon_s,\ \delta = \delta_e + \delta_s)$-差分プライバシを満たすが，厳密ではないため，*RDP* を用いてより厳密な合成を行う．各構成要素で *RDP* を考える．*DP-PCA* は ϵ_p-差分プライバシを満たすため，*[9]* の *Lemma 1* より，$(\alpha, \epsilon_{rp}(\alpha) = 2\alpha\epsilon_p^2)$-*RDP* を満たす．次に，*DP-SGD* では，*DP-SGD* のモーメントに関する *Lemma 3 [5]* と *RDP* の定義より，次の $\epsilon_{rs}(\alpha) = \frac{1}{\alpha-1}\left(\frac{q^2\alpha(\alpha+1)}{(1-q)s_s^2} + O(q^3\alpha^3/s_s^3)\right)$ に対して，$(\alpha, \epsilon_{rs}(\alpha))$-*RDP* を満たす．ここで，$q$ はサンプリング確率（*DP-SGD* におけるバッチサイズ/データ数）であり，s_s は雑音の大きさである．最後に，*DP-EM* では正規分布に従う雑音を用いるため，混合正規分布に対する *EM* アルゴリズムにおけるコンポーネント数を k，繰り返し回数を j，雑音の大きさを s_e とすると $(\alpha, \epsilon_{re}(\alpha) = j(2k+1)\alpha/2s_e^2)$-*RDP* を満たす．*RDP* の合成定理（定理 *1*）より，*P3GM* は $(\alpha, \epsilon_{rp}(\alpha) + \epsilon_{rs}(\alpha) + \epsilon_{re}(\alpha))$-*RDP* を満たす．*RDP* は *DP* に変換が可能であるため（定理 *2*），以下の定理が成り立つ．

定理 3. *P3GM* は任意の $0 < \delta < 1$，$\alpha > 1$ に対して，$\epsilon = \epsilon_{rp}(\alpha) + \epsilon_{rs}(\alpha) + \epsilon_{re}(\alpha)) + \frac{\log 1/\delta}{\alpha - 1}$ で (ϵ, δ)-差分プライバシを満たす．

4 実　　験

実験では主に生成データが本来のデータセットと同じような特徴を持っていることを示す．そのために，視覚化と機械学習精度の二つの方法による実験を行なった．なお，基本的に (ϵ,δ) の値は一般的に用いられている値 $(1,10^{-5})$ に設定した．

4.1 生成データの視覚化

まず，初めに *MNIST* データセットを用いて実験を行なった．*MNIST* データセットは *784* 次元の *1* から *9* の手書き数字画像 *70000* 枚である．*MNIST* データセットを *P3GM* で $(\epsilon = 1, \delta = 10^{-5})$-差分プライバシを満たすように学習し，ランダムに生成したデータを図 *1* に表示した．比較のために，*VAE*（プライバシ保護なし），*DP-VAE*（*DP-SGD* によるプライバシ保護），*DP-GM [6]* を表示した．*P3GM* の結果は *DP-VAE* よりも雑音が少なく，*DP-GM* よりも多様なデータを生成できており，本来の *VAE* に近い結果となっている．これは，分離した学習工程により，*DP-SGD* がうまく機能したことが要因にあると考えられる．

4.2 機械学習精度による評価

P3GM の目的はデータセットをプライバシを保護しつつ公開することであった．その場合，生成データは元のデータの特徴を捉えられていると同時に，機械学習モデルの学習に用いることができるべきである．例えば，生成したデータを用いてクラス分類機械学習モデルを学習し，そのモデルが実際のデータに対して良い精度でクラス分類ができる合成データは質が良いと言える．このことを示すために，合成データで機械学習モデルを学習し，その精度を実データで検証する実験を行なった．用いたデータセットは六種類で詳しい情報は表 *1* にまとめた．

評価はデータセットを十分割の交差検証を用いて行なった．まず，訓練用データを用いて *P3GM* を学習する．このときラベルを *one-hot vector* に変換し，データと結合することでラベル情報も一緒に生成するように学習する．そして，訓練用データと同じ数のデータをラベルが訓練用データと同じ割合になるように生成する．この合成データを用いてクラス分類

データセット名	データ数	列数	クラス数	正クラス割合 (%)
Kaggle Credit [14]	284807	29	2	0.2
Adult *	45222	15	2	24.1
UCI ISOLET *	7797	617	2	19.2
UCI ESR *	11500	179	2	20.0
MNIST	70000	784	10	-
Fashion-MNIST	70000	784	10	-

表 1: データセット

	AUC-ROC			AUPRC		
	VAE	PGM	**P3GM**	VAE	PGM	**P3GM**
線形回帰	0.9617	0.9454	0.9380	0.6542	0.6865	0.6530
AdaBoost [15]	0.9599	0.9330	0.9146	0.5737	0.6528	0.4574
GBM [16]	0.9619	0.9442	0.9221	0.6838	0.6734	0.5231
XgBoost [17]	0.9395	0.9321	0.9247	0.2745	0.6469	0.4824

表 2: 四種類のクラス分類機械学習モデルの AUC-ROC と AUPRC による評価．VAE と PGM はプライバシ保護をしていない．P3GM は $(1, 10^{-5})$-差分プライバシを満たす

機械学習モデルを学習し，検証用データを用いてクラス分類の精度を検証する．機械学習モデルは表 *2* にあるように線形回帰，*AdaBoost [15]*，*GBM [16]*，*XgBoost [17]* を用いた．二値分類の評価には *area under the receiver operating operating characteristic curve*（*AUC-ROC*）と *area under the precision recall curve*（*AUPRC*）を用いた．なお，*AUC-ROC* はデータのラベルに偏りがある場合，その偏りによって正しい評価ができないことがあるため，影響を受けにくい *AUPRC* を同時に用いている．*AUC-ROC* はテストデータをいかに分類できているかであり，0.5 のときに，全く分類ができてなく，1 のときに完全に分類できている．*AUPRC* はいかに偽陽性を出さずに，真陽性を出すかを評価している．0 のときに真陽性を全く出せず，1 であるときに偽陽性を出さずに真陽性を全て出している．

まず，*P3GM* が差分プライバシを満たさない *VAE* と *PGM* と比較して，どれだけ質が落ちるのかを実験する．表 *2* は *Kaggle Credit* データセットにおける結果で，*VAE* と *PGM*（雑音なし）と *P3GM* を比較している．*VAE* と *PGM* を比較すると，ほぼ結果が変わらないことがわかる．これは *PGM* という確率モデルが *VAE* と同様の表現力を持っていることが示している．次に *P3GM* と比較すると，線形回帰などでは同様のスコアを出しているものの，他のモデルでは低下が見られる．これは，学習に雑音を加えたために，生成データにも雑音が混じったことで，それらのモデルに影響を与えたためであると考えられる．

次に既存の差分プライバシを満たす生成モデルとの比較を行う．表 *3* がその結果であり，$(1, 10^{-5})$-差分プライバシを満たす *DP-GM*，*PrivBayes* と比較している．*Adult* 以外のデータセットで *P3GM* が最高スコアを出していることがわかる．*PrivBayes* はベイジアンネットワークを用いて直接的に依存関係を求めているため，*Adult* のような単純な依存関係で構成されるような場合，スコアが高くなると考えられる．*P3GM* の場合でも，特に *ISOLET* データセットに対しては，*Original* と比べてスコアが低いことがわかる．これは *ISOLET* が次元が高く，多くの学習を要し，さらにデータ数が少ないため，大きな雑音が必要になるためであると考えられる．

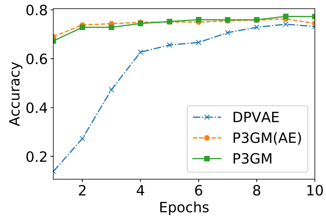


図 3: MNIST データでの各エポックでの正確性.

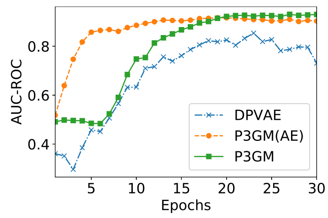


図 4: Kaggle Credit データでの各エポックでの AUC-ROC.

次に，*MNIST*，*Fashion MNIST* データセットで評価を行う．これらはほぼ均一な数のクラスが十クラス存在し，そのクラス分類精度を評価する．同様に，実際のデータセットを訓練用と検証用に分離し，訓練用で生成モデルの学習をし，生成したデータを用いて *NN* によるクラス分類モデルを学習する．スコアは検証用データに対してそのクラス分類モデルを用いたときの正解率とした．表 *4* がその結果である．*P3GM* は，画像データのような高次元の複雑な依存関係も学習することができるため，高品質なデータが生成できたものと考えられる．*DP-GM* の場合，多様性が小さいため *(*図 *1* 参照*)*，生成データでのクラス分類モデルの学習がうまく機能しなかったものと考えられる．

4.3 収束速度についての実験

3.3 節，*3.2* 節で述べたように，*VAE*，*P3GM*，*AE* の順に解の探索範囲が広いと言える．ここでは，それによる効果を実験的に示す．実験は画像データと表データ一種類ずつを用いて行なった．全て，最終的に $(1, 10^{-5})$-差分プライバシを満たすようにパラメータの調整を行なった．図 *3*，*4* がその結果である．ここでは分散を固定した（*3.2* 節参照）モデルを *AE* と呼ぶ．これを見ると両データの結果共に，収束は *AE* が一番速いことがわかる．これは，*AE* が一番探索範囲が狭いためであろう．そして，図 *4* を見ると，*VAE* の学習が安定していないことがわかる．それに対して *AE* と *P3GM* は比較的安定していることがわかる．これはパラメータを固定したことによるものだと考えられる．そして，*P3GM* が *AE* より最終的な結果が良いのは，解の探索範囲が *P3GM* の方が広いことからより良い解に収束できたものと考えられる．しかしその反面，収束に学習回数を要している．よって，精度を捨ててプライバシ保護を強めたい場合は *AE* を用いるべきであると考えられる．

4.4 プライバシ合成についての実験

この論文では *RDP* を用いて *P3GM* の差分プライバシにおける ϵ の値を計算した．ここでは，それが単純な方法より厳密であることを実験的に示す．*Park* ら *[11]* は *DP-EM* において

* : https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition
* : https://archive.ics.uci.edu/ml/datasets/isolet
* : https://archive.ics.uci.edu/ml/datasets/adult

Datasets	AUROC				AUPRC			
	DP-GM	PrivBayes	**P3GM**	Original	DPGM	PrivBayes	**P3GM**	Original
Kaggle Credit	0.8805	0.5520	**0.9232**	0.9663	0.3301	0.2503	**0.5208**	0.8927
UCI ESR	0.4911	0.5377	**0.8243**	0.8698	0.3311	0.4265	**0.7559**	0.8098
Adult	0.7806	**0.8530**	0.8321	0.9119	0.4502	**0.6374**	0.5917	0.7844
UCI ISOLET	0.4695	0.5100	**0.6855**	0.9891	0.1816	0.2099	**0.3287**	0.9623

表 3: P3GM と既存手法の DPGM，PrivBayes との比較．四種類のクラス分類学習モデルを合成データで学習したときの実データでのスコア．Original は合成データではなく，実際のデータセットを用いた場合のスコアである．太字が最高スコアである．

Dataset	VAE	DP-GM	PrivBayes	**P3GM**
MNIST	0.8571	0.4973	0.0970	**0.7946**
Fashion	0.7854	0.5200	0.0996	**0.7311**

表 4: (Fashion) MNIST データセットの合成データで学習したモデルの実データでのクラス分類精度の比較.

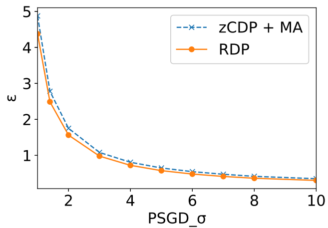


図 5: zCDP + MA（ベースライン）と RDP による差分プライバシの合成の比較.

は，*zCDP* を用いて計算することでより厳密になることを示した．*DP-SGD* では *Moment Accountant (MA) [5]* を用いることでより厳密に計算できることが示されている．よって，単純な方法として，*zCDP* と *MA* をそれぞれ独立に用いることが考えられる．ここでは，*zCDP* と *MA* を用いた場合をベースラインとして，*RDP* を用いた場合を比較する．図 *5* がその結果である．横軸が *DP-SGD* における雑音の大きさを表しており，縦軸が *P3GM* の ϵ の値である．全ての雑音の大きさに対して，ϵ の値が *RDP* を用いた方が小さくなっていることがわかる．

5 結　論

この論文では，*VAE* の学習を簡略化することにより，差分プライバシに必要な雑音がある中でも，より良い解に収束できるような確率モデルを提案した．提案手法の *P3GM* による合成データが既存手法よりも同じプライバシ保護の度合いの時に，より良い質のデータが合成できること実験的にを示した．

文　献

[1] *Cynthia Dwork. Differential privacy.* Encyclopedia of Cryptography and Security, *pp. 338–340, 2011.*

[2] *Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks.* ACM Transactions on Database Systems (TODS), *Vol. 42, No. 4, p. 25, 2017.*

[3] *Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In* Advances in neural information processing systems, *pp. 2672–2680, 2014.*

[4] *Diederik P Kingma and Max Welling. Auto-encoding variational bayes.* arXiv preprint arXiv:1312.6114, *2013.*

[5] *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. ACM, 2016.*

[6] *Gergely Acs, Luca Melis, Claude Castelluccia, and Emiliano De Cristofaro. Differentially private mixture of generative neural networks.* IEEE Transactions on Knowledge and Data Engineering, *Vol. 31, No. 6, pp. 1109–1121, 2018.*

[7] *James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Pate-gan: generating synthetic data with differential privacy guarantees. In* International Conference on Learning Representations, *2018.*

[8] *Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network.* arXiv preprint arXiv:1802.06739, *2018.*

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

[10] *Jack Murtagh and Salil Vadhan. The complexity of computing the optimal composition of differential privacy. In* Theory of Cryptography Conference, *pp. 157–175. Springer, 2016.*

[11] *Mijung Park, Jimmy Foulds, Kamalika Chaudhuri, and Max Welling. Dp-em: differentially private expectation maximization.* arXiv preprint arXiv:1605.06995, *2016.*

[12] *Wuxuan Jiang, Cong Xie, and Zhihua Zhang. Wishart mechanism for differentially private principal components analysis. In* Thirtieth AAAI Conference on Artificial Intelligence, *2016.*

[13] *John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In* 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, *Vol. 4, pp. IV–317. IEEE, 2007.*

[14] *Andrea Dal Pozzolo, Olivier Caelen, Reid A Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. In* 2015 IEEE Symposium Series on Computational Intelligence, *pp. 159–166. IEEE, 2015.*

[15] *Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In* icml, *Vol. 96, pp. 148–156. Citeseer, 1996.*

[16] *Jerome H Friedman. Greedy function approximation: a gradient boosting machine.* Annals of statistics, *pp. 1189–1232, 2001.*

[17] *Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In* Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, *pp. 785–794. ACM, 2016.*