差分プライバシーによるデータ活用最前線

高橋　翼† リュウセンペイ† 長谷川　聡†

† LINE 株式会社　〒 160–0004 東京都新宿区四谷 1-6-1 四谷タワー 23 階
E-mail: †{tsubasa.takahashi,sengpei.liew,satoshi.hasegawa}@linecorp.com

あらまし　差分プライバシーは，グローバルプラットフォーマーや米国の国勢調査等で大規模な統計調査に利用される等，近年注目のプライバシー基準である．差分プライバシーの保証には所定のノイズの加算を用いるため，ノイズの大きさによるプライバシーと有用性のトレードオフの関係がある．一方，差分プライバシーの注目すべき点として，対象とするデータやユーザ群のスケールに応じて有用性の棄損を低減できる点がある．この特性により，ビッグデータを扱う研究や事業において厳密なプライバシーと有用性の両立が期待できることから，差分プライバシーに関する研究は近年大きな進展を見せている．特に，データ合成や転移学習，連合学習といったアプリケーションでは，機微データの実用的な活用を目指した研究開発が加速している．本チュートリアルでは，差分プライバシーの浸透を目的として，初学者向けに配慮した基礎概念の解説，データ合成や連合学習等の差分プライバシーの応用事例の紹介する．

キーワード　差分プライバシー, プライバシー保護機械学習

1 はじめに

データに基づく意思決定やサービスの改善は，経営の合理化や社会情勢の見える化，ユーザ体験や収益の向上に必要不可欠となっている．一方で，ユーザのデータを扱う事業者，特にプラットフォーマーには，ユーザのプライバシーへの配慮が求められている．加えて，国際的な規制の整備も進み，プライバシー保護の技術や考え方もめざましい発展を遂げている．そのため，ユーザのデータを扱うプラットフォーマーには，法令や規制を遵守するだけでなく時流に即した最適なプライバシーモデルの追求と導入が求められている．

差分プライバシー（Differential Privacy, DP）[14] は，統計的なプライバシー基準であり，数学的に厳密なプライバシー保護の水準を示すものである．学術的にはデファクトスタンダードと考えられており，DP を保証したプライバシー保護型のデータサイエンス，機械学習の技術が盛んに研究されている．産業界においても，グローバルプラットフォーマーを中心に，DP を保証したデータ収集やデータ解析が実用化されている．現在，実用化されているものとして，Google Chrome [19] [20] や iOS デバイス [11] からの統計データの収集が知られている．

差分プライバシーを満たすメカニズム (関数や統計処理，クエリ応答，機械学習モデル等) を用いれば，異なる入力によるメカニズムの出力の差異が区別困難であることが統計的に保証される．この区別の困難さは，プライバシー強度 ϵ によって定義し，調整することができる．所定のプライバシー強度 ϵ を保証するためには，メカニズムの特性と ϵ を考慮して作られた特別なノイズの加算が用いられる．一般に，強いプライバシー (小さい ϵ) を保証するためには，大きなノイズが必要となるため，プライバシー強度と出力の有用性との間にトレードオフの関係がある．そのため，プライバシー強度 ϵ の決定は，メカニズムをプライバシー保護する際にクリティカルなパラメータである．しかしながら，ϵ の解釈が直感的ではないこと，適切なノイズ加算方法の設計に知識を要することから，DP を利用するための敷居が高いことが大きな課題である．k-匿名化 [46] のような従来からよく知られたプライバシー基準と比べると，概念が難解で取っ付きづらい印象を持たれている点も普及の阻害要因である．

本チュートリアルでは，差分プライバシーを「やさしく」解説することを目指す．はじめに，できるだけ難解な数式を使わずに，差分プライバシーとは何なのかを初学者でもわかりやすいように配慮して (厳密な定義よりも分かりやすさを優先して) 導入する．その上で，差分プライバシーの基本的な数学的定義や性質を概説する．チュートリアルの後半では，差分プライバシーの保証を前提とする機械学習フレームワークを紹介する．さらに，差分プライバシーによるデータ活用の最前線として，データ合成 (Data Synthesis) や連合学習 (Federated Learning) の研究事例を紹介する．本チュートリアルを通して，差分プライバシーの普及促進を図り，研究の活性化を狙う．

2 やさしい差分プライバシー

差分プライバシー [14] がどういったプライバシーを表すものなのかについて，例を交えて導入する．わかりやすさを優先するため，差分プライバシーの議論で前提とするいくつかの要素を省略した形で，差分プライバシーとはいかなるものなのかのイメージを提供する．

ここでは，リンゴとバナナ，ブドウが好きな Alice と Bob に，「いま一番食べたいフルーツは？」と質問 (クエリ) を投げかけることを考える．Alice と Bob の回答を観察していると，Alice は (リンゴ, バナナ, ブドウ) を (0.65, 0.22, 0.13) の確率で，Bob は (0.68, 0.17, 0.15) の確率で回答する傾向にある，ことが分かった．Alice や Bob が質問に回答するメカニズムがあるとすれば，二人のメカニズムは，似た回答 (出力) をする傾

向にあると言える．このようなメカニズム間の出力の傾向の類似性 (区別の困難さ, 識別困難性) を表す尺度の一つが差分プライバシーである．二つのメカニズムの出力の傾向が似ているなら，出力だけを観測して，出力がどちらのメカニズムに由来するのかを識別することも難しくなる．

次に，上記の二人とは嗜好の異なる Charlie を考える．同様に，Charlie は，(0.00, 0.10, 0.90) の確率で回答する傾向にあることが分かっている．Alice と Charlie は，出力の傾向が大幅に異なるため，出力から二人のメカニズムを容易に区別できる．

Alice と Charlie，二人のメカニズムを区別が難しくなるようにするためには，出力の傾向を近づけるようにするための仕組みの導入が求められる．差分プライバシーでは，ノイズの加算 (ランダム化) によってこれを達成できることが知られている．ノイズの加算の結果として，Alice のランダム化メカニズムは (0.45, 0.34, 0.21)，Charlie のランダム化メカニズムは (0.23, 0.28, 0.49) といった出力確率とすることができ，ランダム化以前よりも出力の傾向を近づけることで区別を難しくできる．

では，どのくらいのプライバシー (区別の難しさ) が保証されているのだろうか？ 差分プライバシーでは，メカニズム間の出力の区別の難しさをプライバシー強度 ϵ で表す．また，ϵ を用いて ϵ-差分プライバシーと呼ぶ．(本来の定義とは若干異なるが)，Alice と Bob のメカニズム $\mathcal{M}_{Alice}$ と $\mathcal{M}_{Bob}$ が ϵ-差分プライバシーであるならば，以下の関係が成り立つ．

$$\Pr(\mathcal{M}_{Alice} = y) \leqq \exp(\epsilon) \Pr(\mathcal{M}_{Bob} = y) \qquad (1)$$

ここで，$y \in \{$ リンゴ, バナナ, ブドウ $\}$ である．ϵ は小さいほどプライバシー強度が高く，0 のときに両辺のメカニズムが全く同じ出力傾向を持つことを表す．式 (1) は，ϵ が小さい (概ね $\epsilon < 1$) とき，以下のように表すこともできる．

$$\Pr(\mathcal{M}_{Alice} = y) \leqq (1 + \epsilon) \Pr(\mathcal{M}_{Bob} = y) \qquad (2)$$

ここまでで述べたストーリーは，わかりやすさを優先したものであり，差分プライバシーの厳密な定義とは異なる点がある．上記の例では，二つの異なるメカニズム間の出力の傾向を比較していた．実際の差分プライバシーでは，あるメカニズムに対して入力が変化したときに，出力の傾向がどう変わるのかを扱う．あくまで，差分プライバシーの沼への招待であり，奥深い引き続き差分プライバシーの世界を覗いていただきたい．

3 差分プライバシーの定義・性質・設計・解釈

本節では，差分プライバシーの定義，重要な概念，基本的な性質について導入する．特に，差分プライバシーを利用する際に考えるべき点を以下の 3 つの要素に分けて述べる．

（1） プライバシーモデル
（2） プライバシー消費の管理
（3） ノイズの設計

また，差分プライバシーを解釈する手段についても紹介する．

3.1 プライバシーモデル

上記の「やさしい差分プライバシー」とは異なり，本来の差分プライバシーは，あるメカニズムにおいて，入力がわずかに異なるときに出力の傾向を区別が難しくすることを考える．

3.1.1 (セントラル) 差分プライバシー

標準的な差分プライバシーでは，この入力がわずかに異なるを表現したものとして，隣接データベースと呼ばれるものを考える．隣接データベースは，あるデータベース D があるとき，D と任意の 1 つのレコードのみが異なるあらゆるデータベース D' である．D と D' には，$d_H(D, D') = 1$ という関係が成り立つ．ここで，$d_H(\cdot, \cdot)$ はデータベース間のハミング距離である．

ここで，D' を D からある 1 つのレコード $x(\in D)$ を削除したものとする ($D' = D \setminus x$)．メカニズムの出力 $f(D)$ と $f(D')$ を観測しても，双方の出力 (の傾向) が区別困難であれば，削除したレコード x の情報を得ることが難しい．これをプライバシー強度 ϵ を用いて示したものが，ϵ-差分プライバシーである．

定義 1 (ϵ-差分プライバシー [14])**.** (ランダム化) メカニズム $\mathcal{M} : \mathcal{D} \to \mathcal{S}$ が ϵ-差分プライバシーを満たすとは，$d_H(D, D') = 1$ を満たす任意の隣接データベースの組 $D, D' \in \mathcal{D}$，および任意の出力の部分集合 $S \subseteq \mathcal{S}$ に対して以下が成り立つときである．

$$\Pr(\mathcal{M}(D) \in S) \leqq \exp(\epsilon) \Pr(\mathcal{M}(D') \in S) \qquad (3)$$

以降，ϵ-差分プライバシーを ϵ-DP と表現する．ϵ は 0 以上の数であり，小さいほどプライバシー強度が高く，0 のときに最も強い．特段 DP を保証していないノンプライベートな状態を $\epsilon = \infty$ で表す．

ϵ-DP を緩和したプライバシー基準として，(ϵ, δ)-DP がある．

定義 2 ((ϵ, δ)-差分プライバシー [14])**.** (ランダム化) メカニズム $\mathcal{M} : \mathcal{D} \to \mathcal{S}$ が $\epsilon \geqq 0$，$\delta \in [0, 1]$ において (ϵ, δ)-差分プライバシーを満たすとは，$d_H(D, D') = 1$ を満たす任意の隣接データベースの組 $D, D' \in \mathcal{D}$，および任意の出力の部分集合 $S \subseteq \mathcal{S}$ に対して以下が成り立つときである．

$$\Pr(\mathcal{M}(D) \in S) \leq \exp(\epsilon) \Pr(\mathcal{M}(D') \in S) + \delta \qquad (4)$$

後述のローカル差分プライバシーとの対比を明確にする際には，上述の差分プライバシーをセントラル差分プライバシー (CDP) と呼ぶことがある．

3.1.2 ローカル差分プライバシー

上述のセントラル DP では，収集済みのデータベースの統計量を第三者に公開する際に，隣接データベースの識別不能性を考えた．一方で，データ所持者 (クライアント) がデータ収集者 (サーバー) にデータを提供する際のプライバシーについては考えられていない．ローカル差分プライバシー (Local DP, LDP) [12] は，データ所持者がデータ収集者にデータを送信する際の識別困難性を扱う．

LDP では，CDP で用いた隣接データベースではなく，任意の入力 $x, x' \in \mathcal{X}$ を対象とする．$\mathcal{X}$ は (入力) データのドメインを表す．

定義 3 (ϵ-ローカル差分プライバシー [12])．(ランダム化) メカニズム $\mathcal{M}: \mathcal{X} \to \mathcal{S}$ が ϵ-ローカル差分プライバシーを満たすとは，任意の入力の組 $x, x' \in \mathcal{X}$，および任意の出力の部分集合 $S \subseteq \mathcal{S}$ に対して以下が成り立つときである．

$$\Pr(\mathcal{M}(x) \in S) \leqq \exp(\epsilon) \Pr(\mathcal{M}(x') \in S) \tag{5}$$

LDP を保証することで，クライアントからサーバーにデータを集める際に，収集されるデータに対して識別困難性を保証できる．そのため，データ収集をするサーバーを信頼する必要がない．LDP は，クライアント群が持つアイテムの頻度推定 [4] [6] [19] [44] [48] や，クライアント群とサーバーとの連合学習 [17] [42] [23] といったユースケースで用いられている．

3.2 プライバシー消費の管理

差分プライバシーでは，データにアクセスする度に「プライバシーを消費する」と考える．これは二度三度とデータを評価した結果を利用する度に，データに関する情報を引き出すことができることに由来する．そのため，データにアクセスすることで消費したプライバシーを合算して，トータルでどの程度のプライバシー消費があったのかを考える必要がある．

定理 1 (直列合成定理 [14])．メカニズム $M_1, \ldots, M_k$ は，それぞれ $(\varepsilon_1, \delta_1)$-，…，$(\varepsilon_k, \delta_k)$-DP を満たすとする．このとき，$M_1, \ldots, M_k$ を順番に適用するようなメカニズムは，$\left(\sum_{i \in [k]} \varepsilon_i, \sum_{i \in [k]} \delta_i\right)$-DP を満たす．

直列合成定理は，差分プライバシーにおける最も単純なプライバー消費の合算方法であり，極めてルーズであることが知られている．本来は繰り返しデータを評価して利用したとしても，プライバシー消費の合計は必ずしも線形には増えるわけではなく，よりタイトな見積もりを導出できる可能性がある．よりタイトなプライバシーの合成の方法として，Rényi Differential Privacy (RDP) [40] による方法や，Advanced Composition [33] による方法などが知られている．

差分プライバシーを適切に利用するためには，センシティビティに基づくノイズの加算方法の設計だけでなく，メカニズムのプライバシー消費を管理することが非常に重要である．メカニズムに許される最大のプライバシー消費量をプライバシーバジェットと呼ぶ．

他方，DP を保証されたメカニズムからの出力は，追加のデータアクセスさえなければ，それ以上にプライバシーを消費しない．これを Post-processing Theorem (後処理定理) と呼ぶ．

3.3 センシティビティとノイズの設計

明示的に ϵ-DP や (ϵ, δ)-DP を満たす状況，すなわち，出力が区別困難である状況を作り出すためには，任意のレコード $x \in D$ が出力に与える影響を制限することが必要になる．これを達成するために，差分プライバシーでは，メカニズムの動作や出力にノイズを加算することで，ランダム性を高めることを考える．このとき，入力の変化が出力に与える影響を制限することを目的とすることから，入力の変化が出力に与える影響度をセンシティビティという概念で表現し，センシティビティとプライバシー強度 ϵ (および δ) で設計されたノイズを加算する．

定義 4 (センシティビティ (Sensitivity))．隣接データベース $\forall D, D'$ に対する関数 f のセンシティビティは以下のように表される．

$$\Delta_f = \sup_{D, D'} ||f(D) - f(D')||_p \tag{6}$$

3.3.1 ラプラスメカニズム

ϵ-DP を保証するメカニズムとして，最もよく知られたものにラプラスメカニズム [14] [15] [16] がある．ラプラスメカニズムでは，平均 0，分散 Δ_f/ϵ で設計されたラプラス分布 $\mathrm{Lap}(0, \Delta_f/\epsilon)$ からノイズをサンプリングし，f の出力に加算する．

$$\mathcal{M}(D) = f(D) + \mathrm{Lap}(0, \Delta_f/\epsilon) \tag{7}$$

同様に，(ϵ, δ)-DP を保証するメカニズムとしてガウスメカニズム [14] [15] [16] が，ϵ-LDP を保証するメカニズムとしてランダムレスポンス [12] [31] による方法が知られている．

3.3.2 DP な学習フレームワーク：DP-SGD

Differentially private stochastic gradient descent (DP-SGD) [1] は，差分プライバシーの下で機械学習モデルを訓練するための汎用的な最適化フレームワークである．DP-SGD のキーアイディアは，訓練時に用いる勾配にノイズを加算することである．ノイズの加算には，前述の通り，センシティビティに応じたノイズの設計が必要になる．勾配のセンシティビティは定義することが難しいため，**クリッピング**という操作を用いることで，勾配のノルムを定数 C で制限する (すなわち勾配のセンシティビティを C に規格化する)．勾配 $\mathbf{g}$ のノルムのクリッピングは以下の計算によって実現される．

$$\pi_C(\mathbf{g}) = \mathbf{g} * \min\left(1, \frac{C}{||\mathbf{g}||_2}\right) \tag{8}$$

DP-SGD では，クリッピングを訓練データ一つ一つに対して実施してから集約し，ガウスノイズを加算することでランダム化された勾配 $\tilde{\mathbf{g}}$ を得る．このとき，ガウスノイズのスケールは，クリッピングサイズ C とノイズスケール σ で決定する．DP-SGD の計算手順をアルゴリズム 1 に示す．DP-SGD による訓練を繰り返した場合のプライバシー合成の方法として，モーメントアカウントが知られている．このプライバシー合成の計算ツールが Tensorflow privacy [1]で公開されている．

3.4 差分プライバシーの解釈

差分プライバシーを統計的仮説検定を用いて解釈するアプローチが提案されている．メカニズム $\mathcal{M}$ の入力 D, D' と出力 y について，以下のような仮説検定を考える．

H_0 : 出力 y は入力 D から作られた．

H_1 : 出力 y は入力 D'から作られた．

1 : https://github.com/tensorflow/privacy

Algorithm 1 DP-SGD

Input: $x_1, \ldots, x_N$

Hyper Parameters: learning rate η_t, noise scale σ'_ϵ, batch size B, clipping size C

Initialize parameters θ_0 randomly

for t **in** $[T]$ **do**

randomly sample *batch* b with probability B/N

$$\tilde{\mathbf{g}} \leftarrow \frac{1}{B}\left(\sum_{x\in b} \pi_C\left(\nabla_{\theta_t}\phi\left(x;\theta_t\right)\right) + \mathcal{N}\left(0,(\sigma C)^2\,\mathbf{I}\right)\right)$$

$$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}$$

end for

Output: θ_T

棄却領域を S，S の補集合を $\bar{S}$ とする．帰無仮説 H_0 が実際には真であるのに棄却した割合 (FP:False Positive Rate) は，$s \in S$ として $\Pr(\mathcal{M}(D) = s)$ と定義される．そして，帰無仮説 H_0 が実際には偽であるのに棄却されなかった割合 (FN:False Negative Rate) は，$\bar{s} \in \bar{S}$ として $\Pr(\mathcal{M}(D') \in \bar{S})$ と定義される．$\mathcal{M}$ が (ϵ, δ)-DP を保証するとは，以下の条件を満たすと同等である [33].

定理 2 (経験的 (ϵ, δ)-差分プライバシー)**.** $\epsilon \in \mathbb{R}^+$, $\delta \in [0, 1]$ について，メカニズム $\mathcal{M}$ は隣接する全てのデータベースの組 D, D' および任意の棄却領域 $S \subseteq Y$ 対して次の条件が満たされる場合にのみ，(ϵ, δ)-DP を満たす．

$$Pr(\mathcal{M}(D) \in S) + e^{\epsilon} Pr(\mathcal{M}(D') \in \bar{S}) \geqq 1 - \delta \tag{9}$$

$$e^{\epsilon} Pr(\mathcal{M}(D) \in S) + Pr(\mathcal{M}(D') \in \bar{S}) \geqq 1 - \delta \tag{10}$$

定理 2 を変形すると，経験的なプライバシー強度 ϵ_{emp} は

$$\epsilon_{emp} = \max\left(\log\frac{1-\delta-FP}{FN}, \log\frac{1-\delta-FN}{FP}\right) \tag{11}$$

と表せる．

上述の仮説検定による経験的な差分プライバシーの考え方に基づいて，DP-SGD の経験的なプライバシ強度を検査する方法が提案されている [41].

また，差分プライバシーを解釈するための別のアプローチとして，差分プライバシーがベイズファクターの観点で，どの程度の証拠能力を持つかについて関係が示されている [28].

4 オープンソース

DP の実装は，理論さえ理解してしまえば難解ではない．一方，初学者には，実際に動かして挙動から理論を理解する機会にも学習効果が期待できる．ここでは，DP に関するオープンソースのライブラリやミドルウェアを紹介する．

DP のライブラリとして，OpenDP[2]や Diffprivlib [27] が公開されている．Google も DP のライブラリを公開している [3]．これらは基礎的な DP を保証する演算をサポートとしている．

2：https://github.com/opendp/opendp

3：https://github.com/google/differential-privacy

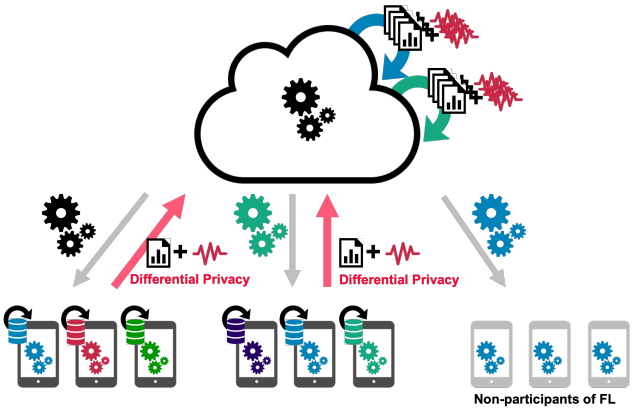

図 1: LDP を保証した勾配の送信による連合学習

DP を保証するデータ解析を試す環境として，PINQ [39], PrivateSQL [36], Flex [29] が提案され，コードが公開させている．Ektelo [53] は，DP を保証する様々なアルゴリズムを組むことための抽象的な部品 (関数群) を提供している．

DP-SGD に関するライブラリとして，TensorFlow Privacy[1] と Opacus[4]がある．Opacus では，ノルムのクリッピングといった DP-SGD のボトルネックとなる処理を高速に処理する工夫がされている．PySyft[5]では，Federated Learning などを Central DP 下で実現する方法がサポートされている．

5 連合学習と差分プライバシー

本チュートリアルでは，LDP を保証する連合学習 (Federated Learning, FL) の実装方法について紹介する．また，LDP 下における実用的な FL を実現するための課題についても言及する．

連合学習 (Federated Learning, FL) は，多数のクライアントと中央の集中サーバーとが共調して機械学習をするフレームワークである [32]．クライアントが保有するローカルデータ自身はクライアント内に留め，ローカルデータを用いた評価値 (主に勾配) のみをサーバーと共有することから，プライバシー保護型の機械学習の一種とも考えられている．しかし，深層学習モデルは，訓練に用いたデータセットに関する機密情報を漏洩する可能性があることが知られている [7] [8] [45]．クライアントが公開する勾配から元の画像が復元可能であることも指摘されている [22] [50].

このようなデータの推定を防ぐ一つの方法としては，勾配に LDP を保証することが挙げられる [17] [42] [23]．図 1 に LDP を保証した勾配の送信による連合学習のイメージ図を示す．

5.1 LDP-SGD

ここでは，勾配をランダム化するアルゴリズムである LDP-SGD(Locally Differentially Private Stochastic Gradient) [13] [17] について述べる．クライアントサイドの計算手順をアルゴリズム 2 に，サーバーサイドの計算手順をアルゴリズ

4：https://opacus.ai/

5：https://github.com/OpenMined/PySyft

Algorithm 2 LDP-SGD; client-side [17]

Require: Local privacy parameter: ϵ_ℓ, current model: $\theta_t \in \mathbb{R}^d$, ℓ_2-clipping norm: L

1: Compute clipped gradient
$x \leftarrow \nabla\ell(\theta_t; d) \cdot \min\left\{1, \dfrac{L}{||\nabla\ell(\theta_t; d)||_2}\right\}$

2: $z_i \leftarrow \begin{cases} L \cdot \dfrac{x}{||x||_2} & w.p. \dfrac{1}{2} + \dfrac{||x||_2}{2L} \\ -L \cdot \dfrac{x}{||x||_2} & otherwise. \end{cases}$

3: Sample $v \sim_u S^d$, the unit sphere in d dimensions.
$\hat{z} \leftarrow \begin{cases} \mathrm{sgn}(\langle \mathrm{z}, \mathrm{v} \rangle) \cdot \mathrm{v} & w.p. \dfrac{e^{\epsilon_\ell}}{1 + e^{\epsilon_\ell}} \\ -\mathrm{sgn}(\langle \mathrm{z}, \mathrm{v} \rangle) \cdot \mathrm{v} & otherwise. \end{cases}$

4: **return** $\hat{z}$

Algorithm 3 LDP-SGD; server-side [17]

Require: Local privacy budget per epoch: ϵ_ℓ, number of epochs: T, parameter set: C

1: $\theta_0 \leftarrow \{0\}^d$

2: **for** $t \in [T]$ **do**

3: Send θ_t to all clients

4: Collect shuffled responses $(\hat{z}_i)_{i \in [n]}$

5: Noisy gradient: $g_t \leftarrow \dfrac{L\sqrt{\pi}}{2} \cdot \dfrac{\Gamma(\frac{d-1}{2} + 1)}{\Gamma(\frac{d}{2} + 1)} \cdot \dfrac{e^{\epsilon_\ell} + 1}{e^{\epsilon_\ell} - 1} \left(\dfrac{1}{n} \sum_{i \in [n]} \hat{z}_i \right)$

6: Update: $\theta_{t+1} \leftarrow \prod_C (\theta_t - \eta_t \cdot g_t)$, where $\prod_C(\cdot)$ is the ℓ_2-projection onto set C, and $\eta_t = \frac{||C||_2 \sqrt{n}}{L\sqrt{d}} \cdot \frac{e^{\epsilon_\ell} - 1}{e^{\epsilon_\ell} + 1}$

7: **end for**

8: **return** $\theta_{priv} \leftarrow \theta_T$

ム 3 に示す．直感的には，勾配は図 2 のようにコインの表が出れば緑色の範囲からサンプリングされ，裏が出れば白色の範囲からサンプリングされる．各クライアントは，まず勾配のノルムが最大でも L になるようにクリッピングする．次にノルムに比例して表が出やすくなるコインを投げ，裏が出た場合には勾配の符号を逆転させる．最後に，ϵ に比例して表が出る確率が高くなるコインを投げ，表が出た場合には元の勾配との内積が正になるように一様分布からサンプリングし，裏が出た場合には内積が負になるように一様分布からサンプリングする．このようにしてランダム化された勾配をサーバーが受け取り，モデルを更新することで連合学習を行う．

5.2 LDP-SGD による連合学習

LDP-SGD を用いた連合学習は以下のフローで実現できる．

（1） サーバーは，クライアント群の一部にグローバルモデルを配布

（2） グローバルモデルを受け取ったクライアントは，保有するローカルデータを用いて勾配を計算

（3） クライアントは，LDP-SGD で勾配をランダム化

（4） クライアントは，所定のタイミングでサーバーにランダム化した勾配を送信

（5） サーバーは，クライアント群から受け取ったランダム化された勾配を平均化してグローバルモデルを更新

（6） (1) ∼ (5) を参加するクライアント群を変えて繰り返す

LDP-SGD でランダム化された勾配は，それら一つ一つでは全く意味をなさないほどにランダム化されている．そのため，ある程度大きな数のランダム化された勾配を平均化することで，ランダム性の打ち消し (ノイズキャンセリング) が必要がある．十分なノイズキャンセリングを実現するためには，訓練するタスクやモデルの複雑性にもよるが，数千から数十万といった規模で平均化を実施する必要がある．また，プライバシーバジェットの観点から各クライアントが 1 回ないし数回しか送信できないことを鑑みると，連合学習に参加するクライアントの数は膨大であることが前提となる．

LDP においては，実用的なプライバシー基準 (ϵ が 1~2 程度) の達成に，非常に大きなノイズの加算が必要であることが知ら

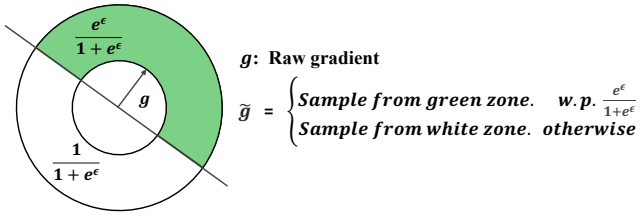

図 2: LDP-SGD による勾配のランダム化

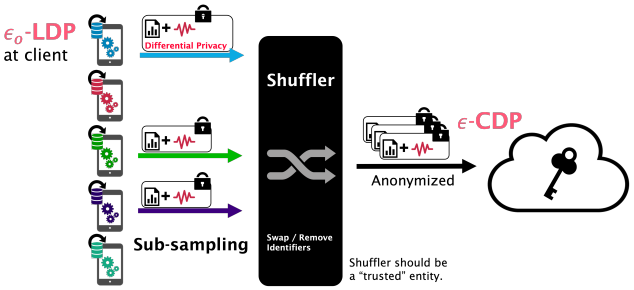

図 3: Shuffling と Subsampling によるプライバシー増幅

れている．CDP との比較においても，データベースへの 1 レコードの追加や削除による出力の差異を識別困難にする CDP と，クライアント毎に異なるあらゆる出力の差異を識別困難にする LDP とでは，LDP の方が問題が難しく，必要なノイズも大きい．LDP の特性として，ユーザ群の規模に応じたユーティリティの向上があるが，それでもなお LDP 制約下では十分な精度のモデルの連合学習は容易ではない．

後述のシャッフルモデルを用いることで，実用的なプライバシー強度を達成可能なことが昨今の研究で明かされている [17]．ただし，連合学習におけるシャッフルモデルの実用的なアーキテクチャはまだ研究が始まったばかりである．

6 近年の研究

最後に，差分プライバシーに関連する最新の研究事例として，三つの研究トピックを紹介する．

6.1 シャッフルモデルによるプライバシー増幅

シャッフルモデル [18] は，CDP と LDP の中間的な位置付

けであり，サーバーとクライアントの間に，シャッフラーと呼ばれる信頼できる第三者を設置する．シャッフラーは，クライアント群から受け取ったデータのアイデンティティを匿名化する，**シャッフリング**と呼ばれる操作を実施する．

（1） クライアントは，LDP を保証したランダム化データを作成し，サーバーの公開鍵で暗号化してシャッフラーに送信

（2） シャッフラーは，(全ての) クライアント群から暗号化されたランダム化データを受け取ったら，それらのアイデンティティ (IP アドレスや送信者 ID など) を付け替える，または削除することにより，クライアント群から受け取ったデータ群に一定の匿名性を保証する．このとき，クライアントから受け取ったデータは，サーバーの公開鍵で暗号化されているためシャッフラーは復号することができない

（3） 最後に，サーバーはシャッフラーから匿名化されたランダム化データ群を受け取る

シャッフリングを用いることで，クライアントで保証された ϵ_0-LDP は，シャッフラーからの出力以降は ϵ-CDP として扱うことができる．このとき，シャッフリングに参加するクライアントの数に応じて，$\epsilon \leqq \epsilon_0$ となるようなプライバシー増幅の効果があることが知られている．

シャッフルモデルと同様にプライバシー増幅を狙ったアーキテクチャが提案されている．クライアントのサブサンプリングによるプライバシー増幅の仕組みが，連合学習での活用を目的として提案されている [34]．プライバシー増幅を成立させるためには，サブサンプリングされたユーザーの身元を隠すために，サーバーが信頼できる必要がある．シャッフラーを前提とするサブサンプリングを用いた連合学習を図 3 に示す．ランダムチェックイン [3] と呼ばれる技術は，連合学習環境におけるより実用的な分散プロトコルとして考案されたものであり，中央集権的な信頼できるサーバーによる指揮のもと，ユーザがランダムにチェックインすることによって，プライバシー増幅の効果を狙っている．

現在のシャッフルモデルによるプライバシー増幅の度合いの導出法は，ルーズであることが知られており，日進月歩で発展し，よりタイトな方法が発見されている [2] [9] [18] [21] [25]．

他方，シャッフルモデルの信頼性を高めるアプローチとして，セキュアなハードウェア (TEE) 上でシャッフリング処理を実施する方法についても提案されている [5]．中央集権的なシャッフラーを用いない分散型の手法も提案されている [10]．

6.2 大規模モデルのファインチューニング

DP-SGD を用いることで，勾配による機械学習の際に DP を保証することができる．しかし，DP-SGD で実用的なプライバシー強度を達成するためには，以下の制限がある．

- データへのアクセス回数 (訓練回数) に制限がある
- ノイズの加算により学習効率が下がる

この制限は，大規模モデルであるほど顕著であるため，GPT-3 のような近年発展が目覚ましい大規模汎用モデルにとっては致命的である．大規模言語モデルに対しても，訓練に用いたデータセットに関する機密情報が漏洩するリスクがあることが報告されている [8]．

これらの課題を解決するアプローチとして，プライバシーや機密性の懸念のないパブリックデータで事前学習を行い，機微データに DP-SGD 等によるファインチューニングをする手法が提案されている [52] [51]．言語モデルを例にとると，汎用的な応答 (対話や要約，補完) をする言語モデルはパブリックデータで訓練してしまい，法務文書などの機密データを使った応答のみを機微データを用いてドメイン適応する，ということに相当する．また，大規模なモデルを差分プライベートに訓練するための効率的な手法についても盛んに研究されている [38] [43]．

6.3 プライバシー保護データ合成

近年，プライバシーを保護しながらデータセットを共有する手段の実現を目指して，プライバシー保護データ合成 (Privacy Preserving Data Synthesis, PPDS) の研究が進んでいる．データ合成は，データセットの統計的な特徴を捉えることで模倣したデータを生成する技術である．GAN (Generative Adversarial Nets) [24] や VAE (Variational Autoencoder) [35] などの深層生成モデルが目覚ましい発展を遂げたことで，PPDS にも注目が集まっている．

著名な PPDS の手法として，PrivBayes [54] がある．PrivBayes は，ベイジアンネットワークによって属性間の依存関係であるグラフを学習し，グラフに基づいて合成データを生成する．一方，DP-SGD を応用することで，GAN や VAE の DP 版を訓練する方法も研究されている．DP-SGD をもととする DP-GAN [49] や PATE-GAN [30] などの手法では，$\epsilon = 10$ 程度の弱いプライバシー強度でしか実用的な合成データを生成できず，$\epsilon = 1$ 付近の実用的なプライバシー強度では，有益なデータを生成できない．これは，上述の DP-SGD による訓練時の制限に起因する課題である．

これらの課題を克服するために, P3GM [47] では, End-to-end のモデルの学習ではなく，二段階の訓練アルゴリズムによって，生成モデルの訓練を簡素化すること手法が提案されている．また，元データへの繰り返しアクセスを必要としない手法として DP-MERF [26] が提案されている．DP-MERF は，まずカーネル関数による埋め込みを差分プライベートに実施し，埋め込みベクトルだけを参照して生成器 (Generator) を訓練する．さらに，DP-MERF の発展として，PEARL [37] という手法が提案されている．PEARL は，DP-MERF の弱点であった訓練の自由度の小ささに起因する表現力の問題を，敵対的な目的関数を導入することで解決した．最新の PPDS の手法を用いれば，ϵ が 1 程度の実用的なプライバシー強度であっても，元データの特徴をある程度維持した合成データを生成することができる．

7 結　　論

本チュートリアルでは，差分プライバシーの浸透を目的として，初学者向けに配慮した基礎概念の解説，データ合成や連合学習等の差分プライバシーの応用事例について紹介した．差分プライバシーは応用が始まったばかりであり，必ずしも理解が

進んでいるとは言い難い．少しでも多くの方に概念を理解いただき，研究の題材として興味を持って頂きたい．また，差分プライバシーを社会実装していく上で，本チュートリアルが学習等の参考となれば幸いである．

文　献

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM, 2016.

[2] B. Balle, J. Bell, A. Gascón, and K. Nissim. The privacy blanket of the shuffle model. In *Annual International Cryptology Conference*, pages 638–667. Springer, 2019.

[3] B. Balle, P. Kairouz, B. McMahan, O. Thakkar, and A. Guha Thakurta. Privacy amplification via random check-ins. *Advances in Neural Information Processing Systems*, 33:4623–4634, 2020.

[4] R. Bassily, K. Nissim, U. Stemmer, and A. Thakurta. Practical locally private heavy hitters. *Journal of Machine Learning Research*, 21(16):1–42, 2020.

[5] A. Bittau, Ú. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnes, and B. Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Proceedings of the 26th Symposium on Operating Systems Principles*, pages 441–459, 2017.

[6] M. Bun, J. Nelson, and U. Stemmer. Heavy hitters and the structure of local privacy. In *Proceedings of the 37th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 435–447, 2018.

[7] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, 2019.

[8] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[9] A. Cheu, A. Smith, J. Ullman, D. Zeber, and M. Zhilyaev. Distributed differential privacy via shuffling. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 375–403. Springer, 2019.

[10] E. Cyffers and A. Bellet. Privacy amplification by decentralization. *arXiv preprint arXiv:2012.05326*, 2020.

[11] A. Differential Privacy Team. Learning with privacy at scale. 2017.

[12] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438. IEEE, 2013.

[13] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

[14] C. Dwork. Differential privacy. In *Proceedings of the 33rd international conference on Automata, Languages and Programming-Volume Part II*, pages 1–12. Springer-Verlag, 2006.

[15] C. Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, 2011.

[16] C. Dwork, A. Roth, et al. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407, 2014.

[17] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, S. Song, K. Talwar, and A. Thakurta. Encode, shuffle, analyze privacy revisited: Formalizations and empirical evaluation. *arXiv preprint arXiv:2001.03618*, 2020.

[18] Ú. Erlingsson, V. Feldman, I. Mironov, A. Raghunathan, K. Talwar, and A. Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.

[19] Ú. Erlingsson, V. Pihur, and A. Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.

[20] G. Fanti, V. Pihur, and Ú. Erlingsson. Building a rappor with the unknown: Privacy-preserving learning of associations and data dictionaries. *Proceedings on Privacy Enhancing Technologies*, 2016(3):41–61, 2016.

[21] V. Feldman, A. McMillan, and K. Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. *arXiv preprint arXiv:2012.12803*, 2020.

[22] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

[23] A. Girgis, D. Data, S. Diggavi, P. Kairouz, and A. T. Suresh. Shuffled model of differential privacy in federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2521–2529. PMLR, 2021.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[25] S. Gopi, Y. T. Lee, and L. Wutschitz. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems*, 34, 2021.

[26] F. Harder, K. Adamczewski, and M. Park. Dp-merf: Differentially private mean embeddings with randomfeatures for practical privacy-preserving data generation. In *International conference on artificial intelligence and statistics*, pages 1819–1827. PMLR, 2021.

[27] N. Holohan, S. Braghin, P. Mac Aonghusa, and K. Levacher. Diffprivlib: the IBM differential privacy library. *ArXiv e-prints*, 1907.02444 [cs.CR], July 2019.

[28] N. Hoshino. A firm foundation for statistical disclosure control. *Japanese Journal of Statistics and Data Science*, 3(2):721–746, 2020.

[29] N. Johnson, J. P. Near, and D. Song. Towards practical differential privacy for sql queries. *Proceedings of the VLDB Endowment*, 11(5):526–539, 2018.

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

[31] P. Kairouz, K. Bonawitz, and D. Ramage. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*, pages 2436–2444. PMLR, 2016.

[32] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[33] P. Kairouz, S. Oh, and P. Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

[34] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. D. Smith. What can we learn privately?

SIAM J. Comput., 40(3):793–826, 2011.

[35] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[36] I. Kotsogiannis, Y. Tao, X. He, M. Fanaeepour, A. Machanavajjhala, M. Hay, and G. Miklau. Privatesql: a differentially private sql query engine. *Proceedings of the VLDB Endowment*, 12(11):1371–1384, 2019.

[37] S. P. Liew, T. Takahashi, and M. Ueno. Pearl: Data synthesis via private embeddings and adversarial reconstruction learning. *arXiv preprint arXiv:2106.04590*, 2021.

[38] Z. Luo, D. J. Wu, E. Adeli, and L. Fei-Fei. Scalable differential privacy with sparse network finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5059–5068, 2021.

[39] F. D. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, pages 19–30, 2009.

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

[41] M. Nasr, S. Songi, A. Thakurta, N. Papemoti, and N. Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 866–882. IEEE, 2021.

[42] H. Ono and T. Takahashi. Locally private distributed reinforcement learning. *arXiv preprint arXiv:2001.11718*, 2020.

[43] N. Papernot, S. Chien, S. Song, A. Thakurta, and U. Erlingsson. Making the shoe fit: Architectures, initializations, and tuning for learning with privacy. 2019.

[44] Z. Qin, Y. Yang, T. Yu, I. Khalil, X. Xiao, and K. Ren. Heavy hitter estimation over set-valued data with local differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 192–203, 2016.

[45] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[46] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[47] S. Takagi, T. Takahashi, Y. Cao, and M. Yoshikawa. P3gm: Private high-dimensional data release via privacy preserving phased generative model. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 169–180. IEEE, 2021.

[48] T. Wang, N. Li, and S. Jha. Locally differentially private heavy hitter identification. *IEEE Transactions on Dependable and Secure Computing*, 2019.

[49] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[50] H. Yin, A. Mallya, A. Vahdat, J. M. Alvarez, J. Kautz, and P. Molchanov. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16337–16346, 2021.

[51] D. Yu, S. Naik, A. Backurs, S. Gopi, H. A. Inan, G. Kamath, J. Kulkarni, Y. T. Lee, A. Manoel, L. Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

[52] D. Yu, H. Zhang, W. Chen, J. Yin, and T.-Y. Liu. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning*, pages 12208–12218. PMLR, 2021.

[53] D. Zhang, R. McKenna, I. Kotsogiannis, M. Hay, A. Machanavajjhala, and G. Miklau. Ektelo: A framework for defining differentially-private computations. In *Proceedings of the 2018 International Conference on Management of Data*, pages 115–130, 2018.

[54] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao. Privbayes: private data release via bayesian networks. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 1423–1434, 2014.