DEIM Forum 2023 5a-3-2

連合学習におけるクライアントサイドでの局所差分プライバシの検証

松本　茉倫† 　高橋　翼†† 　リュウセンペイ†† 　小口　正人†

† お茶の水女子大学　〒 112–8610 東京都文京区大塚 2-1-1
†† 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: †marin@ogl.is.ocha.ac.jp, oguchi@is.ocha.ac.jp, ††{tsubasa.takahashi,sengpei.liew}@linecorp.com

あらまし　クライアントに分散された機微データをプライバシ保護しながら活用し，機械学習モデルを訓練する方法として，局所差分プライバシ (以下 LDP: Local Differential Privacy) を適用した連合学習 (以下 FL: Federated Learning) がある．LDP は，プライバシパラメータ ϵ で表される程度に情報の識別性を困難にすることができる一方で，どういった攻撃に対してどの程度の強度があるのかは未知であり，FL のクライアントにとって理解しやすい説明が必要となる．そこで本研究では，FL で送信する勾配の判別可能性を FL のクライアント自身が検査し，経験的なプライバシ強度を得ることを考える．このとき，2 つの勾配を判別可能な確率が高くなるほどにランダム化手法のプライバシ強度が十分でなく，逆に判別可能な確率が低くなるほどプライバシ強度が高いことを示すことができる．また，この勾配の生成方法についてはアクセスレベル (入力画像を加工できる，勾配を加工できる，など) の異なる 5 種類を提案し，クライアントがプライバシ保護レベルについて理解することを助ける．

キーワード　局所差分プライバシ, 連合学習，経験的プライバシ，仮説検定

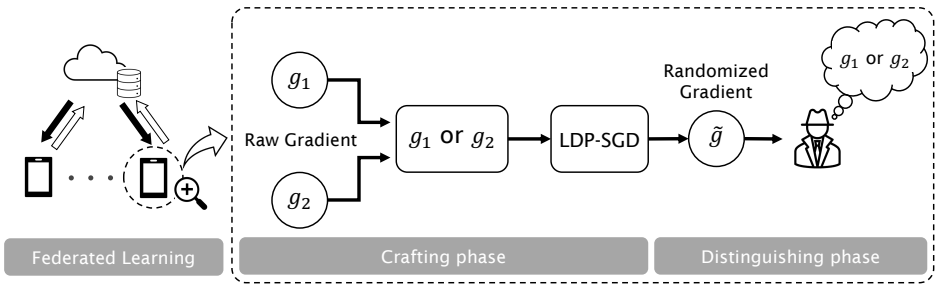
1 はじめに

連合学習 (以下 FL: Federated Learning) [1] はクライアントから生データではなく，勾配をサーバに集約して機械学習を行う手法である．勾配だけをサーバに共有する FL はクライアントのプライバシが保護されるように見えるが，勾配から元画像を復元可能であることを指摘されている [2]．FL において，プライバシ保護した状態で勾配を集める手法の 1 つとして，局所差分プライバシ (以下 LDP: Local Differential Privacy) [3, 4] の適用が挙げられる．まず，差分プライバシ (以下 DP: Differential Privacy) [5] とはプライバシ基準であり，メカニズムが ϵ-DP を満たす場合，メカニズムによる出力を公開したとしても，ϵ 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．標準的な集中型の DP(以下 CDP: Central Differential Privacy) では，信頼できるデータ収集者が正しく DP を満たすメカニズムを使用することを前提としているが，LDP はこの前提を必要としない．そのため，FL におけるクライアントのプライバシ保護に有用であるとされている．LDP を適用したメカニズムはどんな 2 つの入力でもプライバシパラメータ ϵ で表される程度に識別を困難にする．FL の場合，勾配を LDP を適用したメカニズムによってランダム化することで，任意の 2 つの勾配が判別可能な確率は ϵ で表される程度である．

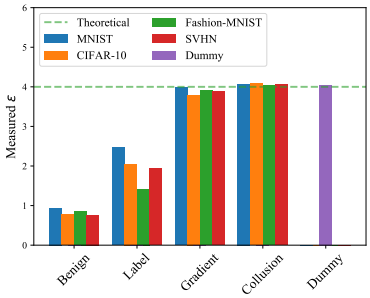
しかしながら，LDP によるプライバシ保護は FL のどんなクライアントにも理解しやすいとは言い難い．LDP を保証したメカニズムは，敵対者が出力から入力の判別に成功する確率の上限 (ワーストケース) を定めることができる．このような説明をサービス提供者から受けた場合，クライアントは納得することができるだろうか．実際に，LDP によるプライバシ保護の説明がユーザの情報共有への意思にどのように影響するか調査した結果 [6] よると，LDP の説明を受けた場合に機微な情報 (生年月日，収入など) を公開するユーザが増加した．一方で，「ランダムなノイズとは何か」,「専門用語が多すぎる」などの回答があり，ユーザは LDP を理解するための助けを必要としていることが示された．LDP が提供するプライバシレベルはプライベート/プライベートではない, のように 2 値ではなくパラメータ ϵ によって制御される統計的なプライバシであり理解が難しい．プライバシレベルの説明は情報共有の意思決定に影響するため，サービス提供者は FL のクライアントに LDP を理解しやすく説明する必要がある．

LDP を保証した FL では，クライアント自身が勾配をランダム化するため信頼できるサーバを必要としない．一方で，クライアントはランダム化メカニズムがどのように勾配をランダム化しているのかを心配する可能性がある．調査 [6] によると，情報共有を許可しなかったユーザはその理由として，DP の技術を信用できないため，アプリケーションまたは企業を信用できないため，と回答している．したがって，より多くのユーザにデータを提供してもらうためには，LDP に関して理解しやすく説明するだけでなく，ランダム化メカニズムが信頼できること示す必要がある．さらに，ユーザ (FL のクライアント) はメカニズムを提供する企業を信用しない場合があるため，FL クライアント自身がメカニズムを検証できる必要がある．

本研究では，FL のクライアント自身で検証可能なプライバシレベルの測定テストを提案する．図 1(a) に示す本研究で提案するプライバシレベルの測定テストは，勾配の生成と LDP を保証した勾配のランダム化を行う Crafting phase とメカニズムの出力から入力を予測する Distinguishing phase に分けられる．この測定テストでは，FL における LDP を保証したランダム化メカニズム LDP-SGD(locally differentially private stochastic gradient descent) [7, 8] を対象として，出力から入



(a) 連合学習のクライアントによる局所差分プライバシの検査



(b) 経験的なプライバシ強度

図 1: (a) 本研究で提案するプライバシ測定テストはメカニズムの入力候補を生成する **Crafting phase** と出力から入力を予測する **Distinguishing phase** で構成され，メカニズムの入力から出力を予測する工程を十分な回数繰り返すとクライアントは経験的なプライバシ強度を得ることができる．このとき，入力の生成方法を変えることでさまざまな攻撃面を実現する．(b) $\epsilon = 4$ を保証したときの経験的なプライバシ強度は，Benign setting において理論値と差があり想定よりも強いプライバシ保護がされていることを示している．また，LDP-SGD のワーストケースである Dummy ではこのメカニズムが ϵ-LDP を満たすことを確認できる．

力 (勾配) を判別可能な確率を算出することで経験的なプライバシ強度 $\epsilon_{\text{empirical}}$ を得る．このとき，さまざまなアクセスレベル (入力画像を加工できる，勾配を加工できる，など) で操作した勾配をメカニズムの入力とすることによって，クライアントがプライバシ保護レベルについて理解することを助ける．

プライバシレベル測定の結果，図 1(b) より以下 4 つの観測が得られた．(1) 悪意を持った勾配の操作を行わない Benign setting において実際に保証しているプライバシレベルは想定よりも強い．(2) 勾配が操作可能な場合は，保証しているプライバシレベル ϵ に達する．(3) FL のサーバとの共謀，すなわち悪意のあるモデルによっては攻撃者の能力を高める．(4) 入力データの正解ラベルを付け加えることは，勾配を操作できない場合でも入力された勾配の判別確率を高める．

本研究の貢献をまとめると以下の 3 つである．

- LDP を保証した FL において，クライアントによる経験的なプライバシ強度の測定テストを提供する．
- LDP-SGD のワーストケース，すなわちメカニズムの出力に最も差が出やすい入力のペアを分析する．
- さまざまな攻撃の場面を想定した経験的なプライバシ強度を測定する．

2 関 連 研 究

2.1 プライバシレベルの説明

LDP は厳密なプライバシ基準であるが，そのプライバシパラメータ ϵ は理解しやすいものではない．文献 [9, 10] では ϵ の代わりの指標を使ったプライバシレベルの説明を提案した．Lee ら [9] はデータベースに発行するクエリの種類やデータの分布・サイズによっては，同じ ϵ を設定してノイズを加えても個人を特定できる確率は異なることを示した．データベース中に個人が含まれる/含まれないを特定できるプライバシリスク ρ は，データベースのレコード数，1 レコードの有無がクエリの出力に与える変化量の最大値すなわち global sensitivity [5]，あるデータセットの 1 レコードの有無がクエリの出力に与える変化量の最大値すなわち local sensitivity [11] から算出される．しかしながら，これらのパラメータはデータ収集時には未知であり，クライアントへの説明としては不向きである．Mehner ら [10] の提案したプライバシリスク $P = 1/(1 + e^{-\epsilon})$ は，データ収集時でも ϵ の説明に利用できる．例えば $\epsilon = 0.1$ を設定した場合は $P = 0.525$ であり，「DP を適用したとき，データベース中にある個人のレコードの有無が特定される確率は最大で 52%」のように説明できる [12]．プライバシリスク P は ϵ を説明することが可能であるが，その説明はワーストケースに限られ，ワーストケース以外の入力に対する出力のプライバシレベルの説明は難しい．本研究では，さまざまな入力に対するプライバシレベルの説明によって，FL のクライアントが LDP を理解することを助ける．

2.2 プライバシ保護の検証

ML Privacy Meter [13] は学習モデルからあるデータが学習データに含まれるかどうかを予想するメンバシップ推定攻撃 [14] によってプライバシリスクを評価する．しかしながら，ML Privacy Meter は DP の検証に特化したものではないため，メカニズムが ϵ-LDP を満たしていることを確認できない．Bullek ら [15] は，Randomized Response と呼ばれる DP の変種を可視化した際のユーザの快適さ，理解，信頼に与える影響を検討した．Bullek らのプロトコルにおいて，ユーザは「過去にドラッグを使用したことがあるか」のようなセンシティブな質問をされたとき，正直に答えるかどうかをルーレットのようなデバイスを使って選択する．この調査では，自身の回答に適用されたランダム化が可視化されることで，ユーザはプライバシ保護メカニズムへの信頼を高めることが示された．Bullek らはルーレットによってメカニズムの可視化を試みているが，本研究で対象とする LDP-SGD [8] は Randomized Response と比べて複雑なメカニズムである．さらに，LDP-SGD が確かに ϵ-LDP を保証していることをユーザ (FL のクライアント) に示すためには，LDP-SGD のワーストケースを分析し，ユーザが再現できる必要がある．

2.3 差分プライバシを保証した機械学習の監査

いくつかの研究 [16–18] では，CDP を適用した機械学習における具体的な攻撃モデルの導入を行った．Liu ら [17] は，仮説検定による CDP の解釈を提案し，Jagielski ら [16] は仮説検定を用いた経験的なプライバシ強度の測定を最初に試みた．彼らは CDP を保証した機械学習アルゴリズムである DP-SGD [19] でプライバシ保護された学習モデルがメンバシップ推定攻撃 [20] と 2 つのポイズニング攻撃 [16,21] に晒された場合のプライバシ強度を測定した．Nasr ら [18] は，汚染されたデータベースを学習に使用した場合，理論的・経験的な ϵ がタイト，すなわちワーストケースであることを示した．逆に，最終的な学習済みモデルのみを見ることが許された場合のように，より制限された敵対者を想定した場合，理論値よりも強いプライバシが保護されている可能性があることを経験的に示した．メンバシップ推定攻撃の代わりに，Bernau ら [22] は，任意の補助知識を持つ敵対者を考え，敵対者が学習データのメンバを識別する際の確実性を制限する (ϵ, δ) に関する識別可能境界として最大ベイズ事後確率の導出を提案した．

3 準　　備

3.1 ϵ-局所差分プライバシ

DP [5] はデータ所持者がデータベース D の統計量を公開する際に，隣接データベース (D と 1 レコードのみ異なるデータベース) の識別不能性を保証することで個人のプライバシを保護する．このとき，信頼できるデータ所持者が正しく DP を満たすメカニズム，つまり信頼できる第三者を必要とする．一方で LDP は，そのような前提を必要とせず，データ提供者が提供の前に自身のデータにノイズを加えることでプライバシを保護する．この場合，各個人が 1 つのデータで構成されるデータベースを所持しており，データそのものという統計量を公開すると捉えることもできる．その場合，隣接データベースはドメイン上の任意のデータであり，DP と同様に任意の隣接データベースとの識別不能性を保証することでプライバシは保護できると考える．$\epsilon \in \mathbb{R}^+$ について LDP は以下のように定義される．

定義 1 (ϵ-局所差分プライバシ)．$x, x' \in \mathcal{X}$ および，任意の出力 $S \in \mathcal{S}$ についてランダム化メカニズム $\mathcal{M} : \mathcal{X} \to \mathcal{S}$ が以下を満たしているとき，$\mathcal{M}$ は ϵ-局所差分プライバシを満たす．

$$\Pr(\mathcal{M}(x) \in S) \leqq e^{\epsilon} \cdot \Pr(\mathcal{M}(x') \in S) \qquad (1)$$

直感的にはメカニズム $\mathcal{M}$ に x を入力とした場合の出力が，任意のデータ x' を入力とした場合の出力と識別することができないため，本来のデータが何であったかが推測できないことを確率的に保証している．

3.2 連 合 学 習

連合学習 (FL) [1] は分散型の機械学習手法である．従来の機械学習と FL の大きな違いはクライアントのデータがサーバや他のクライアントに共有されない点である．本研究の FL では以下のプロトコルに従うものとする．

（1） サーバが n 人のクライアントにグローバルモデル θ_t を配布．

（2） 各クライアントは学習後の勾配 $\nabla\ell(\theta_t; x_i)$ を生成しサーバに送信．

（3） サーバは FedSGD [1] によってクライアントの勾配を集約してグローバルモデルを更新．ここで，η は学習率とする．

$$\theta_{t+1} \leftarrow \underbrace{\theta_t - \eta \cdot \frac{1}{n} \sum_{i=1}^{n}}_{\text{server}} \underbrace{\nabla\ell(\theta_t; x_i)}_{\text{clients}}$$

FL では勾配だけをサーバに共有するが，勾配から学習データを復元できることが指摘されており [2]，プライバシが保護されているとは言い難い．したがって，FL だけではクライアントのプライバシを保護していることにはならないため，何らかのプライバシ保護を取り入れる必要がある．勾配から元データの復元を防ぎ，プライバシを保護する方法の 1 つとして，LDP を適用した勾配のランダム化が挙げられる．

3.3 LDP-SGD

LDP-SGD(Locally Differentially Private Stochastic Gradient Descent) [7,8] は分散環境で動作するように設計されており，クライアントからサーバへの通信全体を最小化することが可能である．本研究で想定する LDP を保証した FL は，信頼されないサーバと機密データを所有するクライアントで構成される．まず，クライアントはサーバから配布されたパラメータを用いて勾配を生成する．次に，クライアントはアルゴリズム 1 のように LDP を保証して勾配をランダム化し，サーバに送信する．サーバは，クライアントから収集した勾配を使い，アルゴリズム 2 によってパラメータを更新する．クライアント側のアルゴリズムは，2 行目と 3 行目で 2 つのランダム化を実行する．ここでは，2 行目を Gradient Norm Projection，3 行目を Random Gradient Sampling と呼ぶこととする．これらは以下の特徴を持つ．

- **Gradient Norm Projection**: 勾配のノルムが L 以上の場合は 100%符号を保ち，L より小さくなるほど符号が反転されやすい．この工程によってノルムは L に矯正される．
- **Random Gradient Sampling**: プライバシパラメータ ϵ を大きくすると，サンプリング前の勾配に近い勾配が生成されやすい．このサンプリングによって生成される $\hat{z}$ は図 2 より以下のように場合分けされる．

$$\hat{z} = \begin{cases} \text{緑の領域からサンプリングする．} & \text{w.p. } \frac{e^{\epsilon}}{1+e^{\epsilon}} \\ \text{白の領域からサンプリングする．} & \text{w.p. } \frac{1}{1+e^{\epsilon}} \end{cases}$$

4 ワーストケースの分析

本節では LDP-SGD のワーストケース，つまり最もメカニズムの出力に差が出やすい入力のペアを分析する．これは，メカニズムが主張するプライバシ強度 ϵ-LDP を確かに満たしているかを確認するために必要である．

Algorithm 1 LDP-SGD; client-side $\mathcal{A}_{\text{client}}$ [8]

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

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

$$z \leftarrow \begin{cases} L \cdot \frac{x}{||x||_2} & \text{w.p. } \frac{1}{2} + \frac{||x||_2}{2L} \\ -L \cdot \frac{x}{||x||_2} & \text{otherwise.} \end{cases}$$

Sample $v \sim_u S^d$, the unit sphere in d dims

$$\hat{z} \leftarrow \begin{cases} \text{sgn}(\langle z, v\rangle) \cdot v & \text{w.p. } \frac{e^\epsilon}{1+e^\epsilon} \\ -\text{sgn}(\langle z, v\rangle) \cdot v & \text{otherwise.} \end{cases}$$

return $\hat{z}$

Algorithm 2 LDP-SGD; server-side $\mathcal{A}_{\text{server}}$ [8]

Require: Local privacy budget: ϵ, number of epochs: T, parameter set: C

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

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

  Send θ_t to all clients

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

  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 $\eta_t = \frac{||C||_2\sqrt{\pi}}{L\sqrt{d}} \cdot \frac{e^\epsilon-1}{e^\epsilon+1}$

end for

return $\theta_{priv} \leftarrow \theta_T$

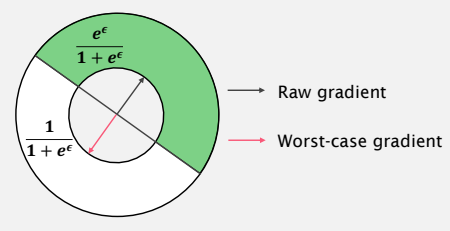


図 2: LDP-SGD によるランダム化．ϵ が大きいほど緑色の領域からサンプリングされやすい．ワーストケースとなるペアはノルムが L 以上である勾配とその勾配と逆向きの勾配である．

4.1 LDP-SGD のワーストケース

アルゴリズム 1 より，LDP-SGD の特徴として，ランダム化前のデータに依存するオブジェクトは，$\hat{z}$ の計算における内積の符号のみであることが挙げられる．すなわち，$0 \ll \lambda$ として，ノルムが大きく異なる勾配のペア $g_1 = (0, 0, ..., 0), g_2 = (\lambda, \lambda, ..., \lambda)$ をメカニズムの入力としても，Gradient Norm Projection によってどちらもノルムが L に矯正されるため判別しやすくならない．さらに，$g_1 = (0, 0, ..., 0), g_2 = (0, \lambda, ..., 0)$ のように透かし λ を挿入した場合でも Randam Gradient Sampling によって紛れてしまうため区別しにくい．したがって，Gradient Norm Projection と Randam Gradient Sampling の影響を受けない勾配のペアを考慮すると，以下の命題が導かれる．

命題 1. ϵ で与えられる理論上の上限に達するワーストケースの入力のペアは，ノルムが L 以上になる勾配とその符号を反転させた勾配である．

証明 1. Gradient Norm Projection によって，ノルムが L 未満の場合は勾配の符号が確率的に反転する．これを防ぐため，勾配のノルムは L 以上である必要がある．これより，g をノルムが L の勾配とする．次に，勾配は Randam Gradient Sampling によって図 2 の緑色の領域から $\frac{e^\epsilon}{1+e^\epsilon}$ の確率でサンプリングされる．2 つの勾配 g_1 と g_2 は反対方向を向いていない限り，ランダム化後の勾配は混ざってしまうため出力に差が出にくい．g_{flip} を g を反転させた勾配，$\tilde{g}_{\text{flip}}$ と $\tilde{g}$ をランダム化後の勾配とする．$\tilde{g}_{\text{target}}$ を 50%の確率で g または g_{flip} をランダム化した勾配としたとき，ランダム化前の勾配はどちらだったかを予想する．ランダム化前後の勾配のコサイン類似度を比較すると以下の場合に分けられる．

（1） $cos(\tilde{g}_{\text{target}}, g) > 0$

- ランダム化された勾配は g であり，±90 度以上回転しなかった．w.p. $\frac{1}{2} \cdot \frac{e^\epsilon}{1+e^\epsilon}$
- ランダム化された勾配は g_{flip} であり，±90 度以上回転した．w.p. $\frac{1}{2} \cdot \frac{1}{1+e^\epsilon}$

（2） $cos(\tilde{g}_{\text{target}}, g) < 0$

- ランダム化された勾配は g_{flip} であり，±90 度以上回転しなかった．w.p. $\frac{1}{2} \cdot \frac{e^\epsilon}{1+e^\epsilon}$
- ランダム化された勾配は g であり，±90 度以上回転した．w.p. $\frac{1}{2} \cdot \frac{1}{1+e^\epsilon}$

Distinguishing phase で，クライアントは $cos(\tilde{g}_{\text{target}}, g)$ が正の場合に g_{target} が g であると予想すると，その予想は確率 $\frac{1}{2} \cdot \frac{e^\epsilon}{1+e^\epsilon}$ で正しい．同様に，$cos(\tilde{g}_{\text{target}}, g)$ が負の場合，クライアントは g_{target} が g_{flip} であると予想するとその予想は $\frac{1}{2} \cdot \frac{e^\epsilon}{1+e^\epsilon}$ の確率で正しいので，クライアントは g と g_{flip} を確率 $\frac{e^\epsilon}{1+e^\epsilon}$ で区別することができる．よって，ノルムが L 以上かつ勾配の符号を反転させることが最も効果的である． □

4.2 ワーストケースに関する制限

4.1 節で示したワーストケースは LDP-SGD の場合に限ったものであり，例えば，DP-SGD [19] のようにクライアントが勾配にガウシアンノイズを加えるメカニズムを採用する場合は異なる可能性がある．DP-SGD のワーストケースについては，Nasr ら [18] によって示されている．

5 経験的なプライバシ強度の検査

2 節で触れたように，プライバシパラメータ ϵ を別の指標で説明する方法としては以下の 3 つが挙げられる．

（1） sensitivity，ϵ をデータベース中にある個人のレコードの有無が特定される確率に変換する [9, 10]

（2） 最大ベイズ事後確率を識別可能境界とする [22]

（3） 仮説検定によって経験的な ϵ を計算する [18]

本研究では，直感的かつさまざまな入力から ϵ を説明可能な (3) の仮説検定を使った経験的なプライバシレベルの測定を採用す

る．本節では，LDP を適用した FL の経験的なプライバシ強度を検査する方法について説明する．

5.1 仮説検定としての局所差分プライバシ

メカニズム $\mathcal{M}$ の入力 x, x' と出力 y について，以下のような仮説検定を考える．帰無仮説を入力 x，対立仮説を x'，棄却領域を S とする．

H_0 : 出力 y は入力 x から作られた．

H_1 : 出力 y は入力 x'から作られた．

帰無仮説 H_0 が実際には真であるのに棄却した割合 (以下 FPR:False Positive Rate) は $\Pr(\mathcal{M}(x) \in S)$ 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 (以下 FNR:False Negative Rate) は S の補集合を $\bar{S}$ とすると，$\Pr(\mathcal{M}(x') \in \bar{S})$ と定義される．メカニズム $\mathcal{M}$ が ϵ-LDP を保証するとは，以下の条件を満たすと同等である [23].

定理 1 (経験的 ϵ-局所差分プライバシ)**.** $\epsilon \in \mathbb{R}^+$ について，メカニズム $\mathcal{M} : \mathcal{X} \to \mathcal{S}$ は任意の入力のペア $x, x' \in \mathcal{X}$ および任意の棄却領域 $S \in \mathcal{S}$ 対して次の条件が満たされる場合にのみ，ϵ-局所差分プライバシを満たす．

$$\Pr(\mathcal{M}(x) \in S) + e^{\epsilon} \Pr(\mathcal{M}(x') \in \bar{S}) \geqq 1$$
$$e^{\epsilon} \Pr(\mathcal{M}(x) \in S) + \Pr(\mathcal{M}(x') \in \bar{S}) \geqq 1$$

定理 1 を変形すると，経験的なプライバシ強度 $\epsilon_{\text{empirical}}$ は

$$\epsilon_{\text{empirical}} = \max\left(\log \frac{1-\text{FPR}}{\text{FNR}}, \log \frac{1-\text{FNR}}{\text{FPR}}\right) \qquad (2)$$

と表せる．例えば 1000 回の試行で，実際には x から作られた出力 y を x' から作られたと予想した割合 (=FPR) が 0.1，実際には x' から作られた出力 y を x から作られたと予想した割合 (=FNR) が 0.2 だった場合，式 (2) より $\epsilon_{\text{empirical}} \simeq 2.0$ となる．ここで注意しなければならないのは，設定する ϵ の値に比例して試行回数を増やす必要があることである．$\epsilon = 4$ であれば試行回数は 10,000 回で十分だが，$\epsilon = 16$ など大きな値を設定した場合はさらに多くの試行を必要とするのでクライアントの環境によってはこの仮説検定は難しくなる．

5.2 LDP における攻撃モデルの導入

上記の仮説検定に基づくプライバシレベルの測定を実施するため，検査を以下のフェーズで構成する．

- **Crafting phase.** FL のクライアントは 2 つの勾配 g_1, g_2 を生成し，どちらか 1 つをアルゴリズム 1 によってランダム化する．ランダム化した勾配は $\tilde{g}$ とする．
- **Distinguishing phase.** $\tilde{g}$ からランダム化された勾配が g_1, g_2 のどちらだったかを予想する．

アルゴリズム 3 には提案するプライバシ測定検査を示す．勾配の操作と判別をある設定のもとで十分な回数繰り返すことで式 (2) より $\epsilon_{\text{empirical}}$ を得る．

5.3 Distinguishing phase

ワーストケースの分析によると，LDP-SGD では出力から入力を判別する際にランダム化前後の勾配のコサイン類似度を比較することが効果的である．このフェーズでは，$\mathcal{D}(\tilde{g}, g_1, g_2)$ を 2 つのランダム化前の勾配 $\{g_1, g_2\}$ とランダム化後の勾配 $\tilde{g}$ のコサイン類似度を利用してランダム化前の勾配を予想する．

$$guess = \begin{cases} g_1 & cos(\tilde{g}, g_1) \geqq cos(\tilde{g}, g_2) \\ g_2 & \text{otherwise} \end{cases} \qquad (3)$$

Algorithm 3 LDP Test in FL clients

Require: Privacy parameter: ϵ, #trials: K

FP, FN, TP, TN $\leftarrow 0$
for $k \in [K]$ **do**
  The FL server sends θ_t to the client.
  Crafting phase
    $\{g_1, g_2\} \leftarrow \text{Craft}(x_1, x_2, \theta_t)$
    Randomly choose g from $\{g_1, g_2\}$
    $\tilde{g} \leftarrow \mathcal{A}_{\text{client}}(g)$.
    Submit $\tilde{g}$ to the distinguishing phase.
  Distinguishing phase
    guess $\leftarrow \mathcal{D}(\tilde{g}, g_1, g_2)$
  if g is g_1 and guess is g_2 **then**
    FP += 1
  else if g is g_2 and guess is g_1 **then**
    FN += 1
  else if g is g_2 and guess is g_2 **then**
    TP += 1
  else
    TN += 1
  end if
end for
Compute $\epsilon_{\text{empirical}}$ as (2)

5.4 Crafting phase

すべての入力のペアの効果を評価することは原理的には可能だが，計算上困難である．本研究では，アルゴリズム 3 における，勾配を操作する方法 Craft(·) として 5 種類を使用する．ここで構築する敵対者は，ベースラインとして最も現実的な勾配を提供するものと，アクセスレベル (入力データを操作できる，勾配を操作できるなど) に応じて 4.1 節で分析されたワーストケースを達成するための勾配を操作するものに分けられる．

5.4.1 Benign setting

最も現実的な設定として，悪意のある勾配を生成しない場合を考える．クライアントはサーバから配布されたグローバルモデル θ_t を用いて，画像 x_1 と x_2 から勾配 g_1 と g_2 を生成する．

$$g_1 = \nabla\ell(\theta_t; x_1);\ g_2 = \nabla\ell(\theta_t; x_2);$$

この設定によって，悪意のある振る舞いをするエンティティを想定しない場合の経験的なプライバシ強度を測ることができる．

5.4.2 Label flip

ワーストケースの分析によると，g_1 と g_2 が最も見分けやす

くなるのは，ある程度ノルムが大きく，g_1 と g_2 が逆を向いている場合であり，勾配を直接操作するのが最も単純なワーストケースの実現方法である．しかしながら，デバイスのメモリアクセスなどの制限によって勾配を操作できない場合がある．この設定では勾配に直接アクセスすることなく，ワーストケースに近づく攻撃として，入力画像の正解ラベルの付け替えを行う．

$$g_1 = \nabla\ell(\theta_t; x_1, label_1);\ g_2 = \nabla\ell(\theta_t; x_1, label_2);$$

g_1 と g_2 の違いはラベルだけであるが，θ_t が事前に良く訓練されているほどこの 2 つの勾配は逆向きになりやすい．次にそのような操作が効果的である理由を説明する．

命題 2. θ_t が十分に事前学習されている場合，入力データの正解ラベルを変更すると，反対方向の勾配が生成される．

証明 2. 2 値分類を行う NN モデルにおいて損失関数を交差エントロピー誤差 $L = -y\log(p) - (1-y)\log(1-p)$ とする．ここで，y は 0 または 1 のラベル，p は $y=1$ の出力確率で，$p = \frac{e^{z_+}}{e^{z_+}+e^{z_-}}$ と表せる．$z_\pm$ はシグモイド変換前の最上位層からの出力 $z_+ = w_+x + b_+$，$z_- = w_-x + b_-$ で，$w_\pm$ と $b_\pm$ はそれぞれ重みとバイアス，x は前の層のロジットとする．正解ラベルを $y=1$ としたとき，勾配は連鎖律より以下のように表せる．

$$\frac{dL}{dW} = \frac{dL}{dp}\cdot\frac{dp}{dz_+}\cdot\frac{dz_+}{dW} = -\frac{1}{p}\cdot p(1-p)\cdot\frac{dz_+}{dW} = (p-1)\cdot\frac{dz_+}{dW} \tag{4}$$

ここで，W は NN の重みである．ラベルを反転させる，すなわち $y=0$ のとき勾配は，

$$\frac{dL}{dW} = \frac{dL}{dp}\cdot\frac{dp}{dz_-}\cdot\frac{dz_-}{dW} = \frac{1}{1-p}\cdot -p(1-p)\cdot\frac{dz_-}{dW} = -p\cdot\frac{dz_-}{dW} \tag{5}$$

となる．$p \simeq 1$ をモデルが事前によく訓練されていると表現すると，z_+ は大きく，z_- は小さくなる．つまり，$w_+ \approx -w_-$ を満たすように重みが学習されることが予想される．$\frac{dz_+}{dW} = \frac{dz_+}{dx}\frac{dx}{dW} = w_+\frac{dx}{dW}$，$\frac{dz_-}{dW} = \frac{dz_-}{dx}\frac{dx}{dW} = w_-\frac{dx}{dW}$ であるから，モデルが事前によく訓練されている，すなわち $y=1$ で $p\simeq 1$ のときは，式 (4) の $p-1$ と式 (5) の $-p$ は符号が同じである．よって，十分に事前学習された NN の場合，ラベルを反転させると勾配が反対方向になる．$y=0$ の場合も同様である． □

5.4.3 Gradient flip

クライアントによる勾配の操作が可能な場合，LDP-SGD のワーストケースに近づく最も単純な攻撃は，逆向きの勾配の生成である．したがって，g_1 と g_2 を以下のように操作する．

$$g_1 = \nabla\ell(\theta_t; x_1);\ g_2 = -g_1;$$

5.4.4 Collusion

Gradient Norm Projection によって，ノルムが L 未満の勾配は確率的に符号が反転されやすいため，Gradient flip のように勾配のペアを逆向きにしてもワーストケースには到達しない場合が発生する．したがって，この設定では Gradient Norm Projection を防ぐため，ノルムが小さくなりにくいモデル $\tilde{\theta}_t$ を故意に生成する．$\tilde{\theta}_t$ は全て同じ正解ラベルを持つデータでサーバが事前に学習したモデルとし，g_1 は $\tilde{\theta}_t$ の学習に使用した正解ラベルとは別のラベルが付けられたデータから生成する．g_2 は Gradient flip と同様に g_1 を反転させた勾配とする．

$$g_1 = \nabla\ell(\tilde{\theta}_t; x_1);\ g_2 = -g_1;$$

5.4.5 Dummy

最も強力な攻撃として，クライアントがダミーの勾配を生成する設定を考える．この設定では，Collusion のようにサーバと共謀せずにワーストケースを達成できるため，FL のクライアントだけで LDP-SGD が ϵ-LDP を満たすことを検証できる．g_1 は Gradient Norm Projection を防ぐためにノルムが L となるように値が埋められた勾配，Random Gradient Sampling の影響を受けにくくするために g_2 は g_1 を反転させた勾配とする．

$$g_1 = (\lambda, \lambda, ..., \lambda);\ g_2 = -g_1;$$

ここで $\lambda = L/\sqrt{d}$，d は勾配の次元である．

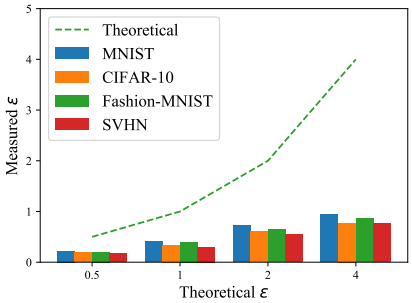


図 3: **Benign setting**: 現実的な設定では，想定されるプライバシレベルよりも勾配は強く保護される．

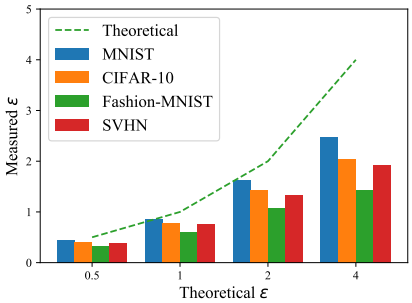


図 4: **Label flip**：モデルが十分に事前学習されている場合，入力データのラベルを変更するだけでも Benign setting より理論値に近づく．

6 実　　験

本節では，FL における LDP のプライバシ測定テストの結果を示す．前述の 5 種類の Crafting phase においてアルゴリズム 3 のようなテストを 10 回行い，得られた $\epsilon_{\text{empirical}}$ を平均する．それぞれのテストの試行回数 K は 10,000 とした．使用したデータセットは MNIST [24]，CIFAR-10 [25]，Fashion-MNIST [26]，SVHN [27] である．

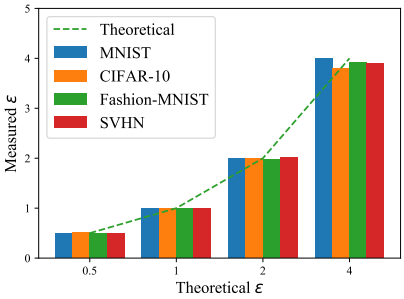


図 5: **Gradient flip**: 勾配のペアの符号が逆向きの場合，経験的なプライバシレベルはほとんど理論値に近づく．

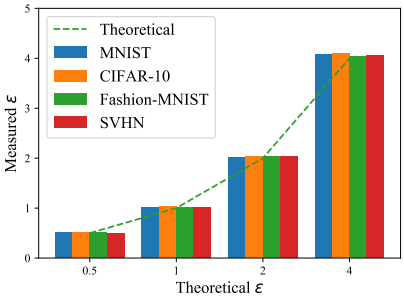


図 6: **Collusion**: サーバと共謀したモデルから符号が逆向きの勾配のペアを生成することで，ワーストケースに到達する．

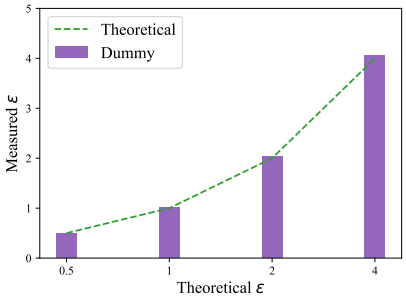


図 7: **Dummy**: クライアントが持つ画像やモデルに関係なくダミーの勾配を生成した場合，LDP-SGD が ϵ-LDP を保証していることを検証できる．

表 1: MNIST における勾配の識別精度．例えば，勾配 g と $-g$ が識別できる確率は，$\epsilon = 1$ を保証した場合，70.5%となる．

	Theoretical ϵ			
	0.5	1	2	4
Benign	54.4	58.4	64.4	68.1
Label flip	60.7	70.1	83.2	92.1
Gradient flip	61.0	70.5	84.7	93.6
Collusion	62.3	73.1	87.9	98.2
Dummy	62.2	73.2	88.2	98.2

6.1 経験的なプライバシ強度の観測

6.1.1 Benign setting

図 3 より，最も現実的な設定では理論値と経験的なプライバシレベルの差が大きい．MNIST で $\epsilon = 4$ を保証した勾配を生成すると，縦軸の経験的な ϵ の値は 0.94 であり，想定よりも強いプライバシ保護がされている．この傾向は全てのデータセットで同様に観測された．

6.1.2 Label flip

図 4 より，入力データを操作することで Benign setting よりもワーストケースに近づく．この設定において，MNIST で $\epsilon = 4$ を保証して勾配を生成すると $\epsilon_{\text{empirical}} = 2.47$ となった．図 4 の結果は，我々の実験環境で限界まで事前学習したモデルを使用した場合であり，学習方法によってはさらにワーストケースに近づく可能性がある．

6.1.3 Gradient flip

単純に勾配を反転させるだけの操作であっても，図 5 のように Benign setting よりもワーストケースに近づくことができる．MNIST で $\epsilon = 4$ を保証した勾配を生成すると $\epsilon_{\text{empirical}} = 3.99$ でほとんど理論値に達しているが，その他のデータセットでは完全には到達していない．

6.1.4 Collusion

図 6 より，クライアントが勾配を直接操作でき，サーバと共謀した結果，全てのデータセットで $\epsilon_{\text{empirical}}$ が理論値に達し，ワーストケースを再現できていることがわかる．

6.1.5 Dummy

図 7 より，クライアントが持っているデータに関わらずダミーの勾配を生成した場合も Collusion と同様に，全てのデータセットで $\epsilon_{\text{empirical}}$ が理論値に達している．すなわち，この攻撃を再現することでクライアントは LDP-SGD が ϵ-LDP を満たしていることを検証できる．

6.1.6 結果のまとめ

図 1(b) と表 1 には 5 つの設定における経験的なプライバシレベルをまとめている．図 1(b) より，ワーストケースを実現するには勾配を直接操作することが有効である．表 1 に示す Crafting phase で操作された 2 つの勾配が Distinguishing phase で判別に成功した確率は，クライアントがプライバシパラメータ ϵ について理解するのを助ける．

6.2 議　　論

6.2.1 クライアントによるワーストケースの実現可能性

実験によって，ワーストケースを実現するにはクライアントが勾配にアクセスできる必要があることが示されたが，デバイスのアクセス制御などの制限によってはワーストケースを再現できない可能性がある．Benign setting では理論値と経験的なプライバシレベルはかけ離れており，Label flip のように入力データを操作可能であればワーストケースに達する可能性があるが，事前学習されたモデルの精度に依存する．

6.2.2 プライバシパラメータの緩和

ワーストケースを想定した攻撃を防ぐ方法として，勾配へのアクセスの禁止が考えられる．実験結果より，勾配を直接操作しない場合は経験的なプライバシレベルは理論値と離れており，LDP を保証した FL のプライバシ保護が過剰で，その実用性を低下させる可能性を示唆している．また，FL のエンティティに何らかの制限を加えることで，ϵ を緩和できる可能性がある．

7 結　　論

FL におけるクライアントのプライバシを保護するため，本研究では LDP を保証した FL を想定し，勾配のランダム化メカニズムとして LDP-SGD を用いた．LDP はプライベート/プライベートではない, のように 2 値ではなくプライバシパラメータ ϵ によって制御されるため，クライアントにとって理解するのが難しく，LDP の明確な説明とメカニズムの信頼性を提供する必要がある．本研究では，クライアントが実行可能な経験的プライバシレベルのテストを提案した．また，メカニズムが ϵ-LDP を満たすかどうかを検証するためのワーストケースを発見し，実験では様々な入力から経験的なプライバシレベルを観測した．最後に，FL のクライアントによるワーストケースの実現可能性と，プライバシパラメータ ϵ の緩和の可能性について議論した．

文　　献

[1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[2] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947, 2020.

[3] Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 211–222, 2003.

[4] Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[5] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[6] Aiping Xiong, Tianhao Wang, Ninghui Li, and Somesh Jha. Towards effective differential privacy communication for users' data sharing decision and comprehension. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 392–410, 2020.

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

[8] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Shuang Song, Kunal Talwar, and Abhradeep Thakurta. Encode, shuffle, analyze privacy revisited: Formalizations and empirical evaluation. *arXiv preprint arXiv:2001.03618*, 2020.

[9] Jaewoo Lee and Chris Clifton. How much is enough? choosing ε for differential privacy. In *International Conference on Information Security*, pages 325–340. Springer, 2011.

[10] Luise Mehner, Saskia Nuñez von Voigt, and Florian Tschorsch. Towards explaining epsilon: A worst-case study of differential privacy risks. In *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 328–331, 2021.

[11] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 75–84, 2007.

[12] Daniel Franzen, Saskia Nuñez von Voigt, Peter Sörries, Florian Tschorsch, and Claudia Müller-Birn. " am i private and if so, how many?"-communicating privacy guarantees of differential privacy with risk communication formats. *arXiv preprint arXiv:2208.10820*, 2022.

[13] Sasi Kumar Murakonda and Reza Shokri. Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. *arXiv preprint arXiv:2007.09339*, 2020.

[14] 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)*, pages 3–18. IEEE, 2017.

[15] Brooke Bullek, Stephanie Garboski, Darakhshan J Mir, and Evan M Peck. Towards understanding differential privacy: When do people trust randomized response technique? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3833–3837, 2017.

[16] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33:22205–22216, 2020.

[17] Changchang Liu, Xi He, Thee Chanyaswad, Shiqiang Wang, and Prateek Mittal. Investigating statistical privacy frameworks from the perspective of hypothesis testing. *Proceedings on Privacy Enhancing Technologies*, 2019:233–254, 07 2019.

[18] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papemoti, and Nicholas Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 866–882. IEEE, 2021.

[19] 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*, pages 308–318, 2016.

[20] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, 2018.

[21] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

[22] Daniel Bernau, Günther Eibl, Philip W Grassal, Hannah Keller, and Florian Kerschbaum. Quantifying identifiability to choose and audit ϵ in differentially private deep learning. *arXiv preprint arXiv:2103.02913*, 2021.

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

[24] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[25] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 2010.

[26] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[27] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.