ポイズニング支援型メンバーシップ推定攻撃に対する差分プライバシーの一考察

芦澤 奈実[1,a)]　矢内 直人[2]

概要：攻撃者が悪意を持って学習データの一部（汚染データ）を提供することで，学習済みモデルにアクセスしてモデルから学習データを効果的に得るデータ復元攻撃が近年に示された．本稿では，学習データが差分プライバシーを満たすことで，上述したデータ復元攻撃のうち，メンバーシップ推定攻撃を防ぐことが可能であると実験的に確認する．また，差分プライバシーの敏感度ごと，および汚染データのサンプル数ごとに，メンバーシップ推定攻撃の成功率を定量的に評価する．これによって，ポイズニング支援型メンバーシップ推定攻撃を防ぐための，差分プライバシーおよびポイズニング攻撃の条件を明らかにする．この目的に向けて，ポイズニング支援型メンバーシップ推定攻撃のフレームワークも示す．実験を行ったところ，50,000 枚の学習データのうち 250 枚の汚染データによって，メンバーシップ推定攻撃の成功率が 25.26% 上昇する結果となった．一方で，$\epsilon \leq 360$ の差分プライバシーを満たすことで，ポイズニング攻撃によるメンバーシップ推定攻撃の成功率の上昇を 0.94% 以内に抑えることに成功した．したがって，差分プライバシーによりポイズニング支援型メンバーシップ推定攻撃を防ぐことを確認した．

キーワード：機械学習, ポイズニング攻撃, メンバーシップ推定攻撃, 差分プライバシー

A Study on Differential Privacy for Poisoning-Assisted Membership Inference Attacks

Nami Ashizawa[1,a)]　Naoto Yanai[2]

Abstract: Recently, model inversion attacks have been proposed in which an attacker maliciously provides a part of the training data (poison data) to access a trained model and effectively obtain the training data from the model. In this paper, we experimentally confirm that differential privacy prevents model inversion attacks, specifically membership inference attacks. We then quantitatively evaluate the success rate of membership inference attacks by differential privacy strength and the number of poison data samples. We determine the parameter of differential privacy and poisoning attacks to succeed in membership inference attacks. To this end, we also present a framework for poisoning-assisted membership inference attacks. We experimentally found that 250 poison data out of 50,000 training data increased the success rate of membership inference attacks by 25.26%. On the other hand, the differential privacy on $\epsilon \leq 360$ reduced the increase in the attack success rate of membership inference attacks to within 0.94%. Therefore, we found that differential privacy prevents poisoning-assisted membership inference attacks.

Keywords: machine learning, poisoning attack, membership inference attack, differential privacy

1. 序論

1.1 背景

近年では自動運転や翻訳，犯罪捜査や防犯カメラの映像

[1] NTT 社会情報研究所, NTT Social Informatics Laboratories
[2] 大阪大学, Osaka University
[a)] nami.ashizawa.ge@hco.ntt.co.jp

解析などをはじめとする様々なアプリケーションで機械学習が利用されている．より高性能なモデルを作るために，クラウド上のデータや顧客等の提供するデータを用いて，あるいは持ち寄って学習を行うことがしばしばある．このとき，収集したデータの中に攻撃者が悪意を持って提供したデータ（汚染データ）が含まれていると，ポイズニング攻撃 [1], [2], [3], [4], [5] によって学習済みモデルの性能が低下する．ポイズニング攻撃は，無差別なクラスを対象にモデルの推論精度を低下させる無差別型 [1], [2]，および特定のクラスを対象にモデルの推論精度を低下させる標的型 [3], [4] が主に提案されている．

また，学習データのプライベートな情報を盗むデータ復元攻撃 [6], [7], [8], [9], [10] の危険性もある．データ復元攻撃は，あるデータが学習データに含まれるか否かを判定するメンバーシップ推定 [6], [7] や，学習データそのものを取り出すデータ抽出 [8], [9] が主に提案されている．

近年，データ復元攻撃の成功率を向上させるために，ポイズニング攻撃を踏み台にしてデータ復元攻撃を行う支援型攻撃が提案されている [11], [12], [13]．このようにポイズニング攻撃を踏み台とする攻撃を，本稿ではポイズニング支援型攻撃と呼ぶ．攻撃者にとっては従来のデータ復元攻撃を行う場合と同じコストで，学習データに関する情報をより多く取得することができる [11], [12].

著者らの知るかぎり，ポイズニング支援型データ復元攻撃への対策は提案されておらず，従来のデータ復元攻撃への有効な対策としては，差分プライバシー [14] が広く知られている．ポイズニング支援型データ復元攻撃について，差分プライバシーでも防ぐことはできないという見解が示されている [11] 一方で，差分プライバシーを満たせば防ぐことが可能であるという考察もされている [12].

1.2 貢献

本稿では，ポイズニング支援型データ復元攻撃のうち，ポイズニング支援型メンバーシップ推定攻撃のフレームワークを提案する．そのうえで実験を通して，差分プライバシーを満たすことで，ポイズニング支援型メンバーシップ推定攻撃を防ぐことが可能であると示す．差分プライバシーの敏感度ごと，および汚染データのサンプル数ごとに，ポイズニング支援型メンバーシップ推定攻撃の成功率を定量的に測定・評価することで，攻撃を防ぐことができる条件を明らかにする．

実験の結果，250 枚以上の汚染データによるポイズニング攻撃を踏み台にすることで，メンバーシップ推定攻撃の成功率が 14.29% 以上上昇した．一方で，$\epsilon \leq 360$ の差分プライバシーによって，メンバーシップ推定攻撃の成功率の上昇を 0.94% 以内に抑えることに成功した．したがって，差分プライバシーを用いてポイズニング支援型メンバーシップ推定攻撃を防ぐことが可能である．

2. 関連研究

2.1 支援型攻撃

支援型攻撃の組み合わせは様々考えられ，モデル抽出攻撃 [15], [16], [17] やポイズニング攻撃 [1], [2], [3], [4], [5] を一段階目として, 二段階目に敵対的サンプル [18], [19], [20] あるいはデータ復元 [11], [12], [21], [22] をなどを用いた攻撃が提案されている．

近年ではポイズニング攻撃を踏み台とするデータ復元攻撃が発見された [11], [12], [13]．ポイズニング攻撃を行うことで，従来のデータ復元攻撃よりも学習データに関するより多くの情報を入手できることが示されている．

ポイズニング支援型データ復元攻撃の成功率上昇について，文献 [12] では，ポイズニング攻撃によって学習データの外れ値を意図的に生み出すことで，データ復元攻撃を容易にすると考えられている．そのため，学習データの外れ値の影響を軽減する効果のある差分プライバシーによって，ポイズニング支援型データ復元攻撃であってもある防ぐことが可能であると見解を示している．

一方で文献 [11] では，ポイズニング支援型データ復元攻撃を，学習データ全体の統計情報等を盗み出す攻撃としている．そのため，個々の学習データのプライバシーを守る差分プライバシーでは，ポイズニング支援型データ復元攻撃を防ぐことはできないと見解を示している．

本稿では，差分プライバシーによってポイズニング支援型データ復元攻撃を防ぐことが可能であるかどうか，実験によって明らかにする．

2.2 ポイズニング攻撃

ポイズニング攻撃は無差別型 [1], [2], [23], [24], [25]，標的型 [3], [4], [26]，バックドア型（トロイとも呼ばれる）[5], [27], [28] に大別される．無差別型は，モデルの精度を下げる．標的型は攻撃対象となるサンプルを定め，そのサンプルに関する推論が誤るように学習させる．バックドア型は，トリガと呼ばれる特殊な入力が与えられたときだけ推論を誤るように学習させる．

近年では，攻撃の検知 [29], [30], [31], [32] を回避するような攻撃もある [33], [34], [35], [36], [37]．本稿の成果は，一段階目となるポイズニング攻撃を上述した文献それぞれの手法に替えた場合であっても，有効性が期待できる．詳細な検討は今後の課題である．

本稿の攻撃者は既存のポイズニング支援型データ復元攻撃 [11], [12] に基づいており，その能力は少量の学習データを追加するだけである．これは，学習アルゴリズムを変更できるような既存の攻撃者 [28], [37] よりも弱い．他にもより強い設定として学習用プログラムを操作できる攻撃者 [38], [39]，モデルのアーキテクチャを指定できる攻撃者 [40], [41] が知られている．本稿の攻撃者は上述した文

献よりも達成しやすい目標を持つことから，これらよりも達成しにくい目標を持つ攻撃者の設定においても本稿の成果は有効である．

2.3 データ復元攻撃

データ復元に関する攻撃はメンバーシップ推定 [6], [7], [42]，属性推定 [10], [43], [44]，データ抽出 [8], [9], [45] に大別される．メンバーシップ推定は，与えられたサンプルが学習データセットに含まれていたか推定する．属性推定は，学習データセットにおいてあるユーザに関する未知の特徴量を，モデルとそのユーザに関する他の情報から推定する．データ抽出は学習サンプルそのものを復元する．

本稿で主に扱う攻撃はメンバーシップ推定 [6] であるが，属性推定やデータ抽出と組み合わせることで，より強力なプライバシーへの脅威を示すことができる．

また，分散設定での学習ではモデルを動的に更新することで，プライバシーに対するより強い攻撃が示されている [46], [47], [48]．本稿の攻撃は，より一般的な学習に対する攻撃とみなせる．

2.4 各攻撃に対する差分プライバシーの位置づけ

ポイズニング攻撃およびデータ復元攻撃それぞれに対して，差分プライバシー [14] が有効であることが示されている．このため，差分プライバシーは多くのライブラリで導入されている [49], [50], [51].

まずポイズニング攻撃に対して, 差分プライバシーによってポイズニングの影響を弱めることが可能である [52], [53]. 他には， 入力を加工する [54]，あるいは入力にノイズを載せた学習を事前と事後に行う [55] ことで，モデルの挙動が変化するような外れ値の影響を防ぐことが可能である．しかし，近年では差分プライバシーに有効なポイズニング攻撃が示されている [56], [57].

一方，データ復元攻撃に対しては，データを識別できなくすることで，攻撃を防ぐことができる [44], [58]．一般的な差分プライバシーの適用では精度への影響が大きいことから，近年ではデータ復元に対してノイズの量を抑える研究が注目されている [59], [60]．しかし，支援型のデータ復元攻撃に対して差分プライバシーの有効性は不明であり，本稿ではそれを明らかにする．

3. ポイズニング支援型メンバーシップ推定攻撃

本節ではポイズニング支援型メンバーシップ推定攻撃を示す．なお，文献 [11], [12] の定義を応用することで，2.3 節で述べたとおり，より強いプライバシー攻撃へも拡張できる．また，差分プライバシーの概念と定義した攻撃の下での，本稿で示す定理を述べる．

3.1 フレームワーク

本稿の攻撃は暗号理論で用いられるゲームベース定義で記載する．まず，データサンプル集合を $\mathcal{X}$，ラベル集合を $\mathcal{Y}$，データセット全体の集合を $\mathcal{D}$，機械学習モデルの集合を $\mathcal{M}$ とする．このとき，機械学習モデル $M \in \mathcal{M}$ は写像 $M : \mathcal{X} \to \mathcal{Y}$ として定義され，モデル M の学習アルゴリズムを写像 $L_M : \mathcal{D} \to \mathcal{M}$ とする．

このとき，攻撃者を $\mathcal{A}$, チャレンジャを $\mathcal{C}$ とすると，ポイズニング支援型メンバーシップ推定攻撃は以下に述べるゲームから定義される．なお，攻撃者 $\mathcal{A}$ はモデル M に対し，クエリを投げることでその出力だけをもらい，他には何も得ないブラックボックス設定とする．ただし，学習アルゴリズム L 及び $\mathcal{X}$ の特徴量は $\mathcal{A}$ にとって既知とする．

(1) $\mathcal{C}$ はデータセット $D \subseteq \mathcal{D}$ を用意する．

(2) ランダムにビット $b \leftarrow \{0, 1\}$ を選択する．$b = 1$ なら D から学習サンプル $z \in D$ を選び，そうでなければ D 以外からランダムにサンプル $z \in \mathcal{D} \backslash D$ を選ぶ．

(3) $\mathcal{A}$ は z を与えられ，m 個のサンプルからなるポイズニング用データセット D_p を用意し，$\mathcal{C}$ に送る．

(4) $\mathcal{C}$ は $M = L_M(D \cup D_p)$ でモデル M を学習する．

(5) $\mathcal{A}$ はデータサンプル $x_1, \cdots, x_q \in \mathcal{X}$ を用いて M から $y_1 = M(x_1), \cdots, y_q = M(x_q)$ を得る．

(6) $\mathcal{A}$ はビット b' を出力する．もし $b = b'$ なら $\mathcal{A}$ はゲームに勝つとする．

Definition 1. 上述したゲームにおいて，攻撃者 $\mathcal{A}$ がゲームに勝つアドバンテージ ϵ を以下のように表すものとする:

$$\epsilon := |Pr[b = b'] - 1/2|$$

このとき，攻撃者は (m, q, ϵ) でメンバーシップ推定攻撃に成功するという．ここで，m は攻撃者が用意するポイズニング用データセットのサンプル数，q はモデル M に行うクエリ回数を意味する．

従来の攻撃 [6], [44] と比べて，上記の攻撃は段階 (3)-(4) がある点が異なる．この部分がポイズニング攻撃に該当する．直観的には，ポイズニング攻撃を通じて攻撃者はモデル M が持つ分布を部分的に制御できる．

3.2 差分プライバシー

差分プライバシー [14] は以下のとおり定義される．

Definition 2. 定義域 D 及び値域 R を持つ秘匿化処理 $f : D \to R$ が，任意の隣接した入力 $d, d' \in D$ 及び出力の集合 $S \subseteq R$ において $\Pr(M(d) \in S) \leq \exp(\epsilon) \Pr(M(d') \in S) + \delta$ が成り立つとき，f は (ϵ, δ)-DP を満たすという．

代表的な実現方法は，f の敏感度に基づいて入力にノイ

ズを与えることである．ここで敏感度とは互いに隣接した任意の入力 d, d' における絶対距離 $|f(d) - f(d')|$ の最大値として定義される．また，ノイズはガウス分布あるいはラプラス分布が良く用いられる．

差分プライバシーを満たすことでメンバーシップ推定攻撃および属性推定攻撃を防ぐことが知られている [44]．ただし，それはポイズニング攻撃（前節のゲームにおける段階 (3)-(4) に相当）を持たない攻撃に対してのみである．2.1 節で述べたとおり，本稿の攻撃に差分プライバシーが有効かは非自明であり，本稿で明かにする．

4. 実験

ポイズニング支援型メンバーシップ推定攻撃 [11], [12] への差分プライバシーの効果を実験で確かめる．

(1) 差分プライバシー満たさず，ポイズニング攻撃を受けていないモデル，(2) 差分プライバシーを満たし，ポイズニング攻撃を受けていないモデルへのメンバーシップ推定攻撃の成功率に対し，(3) 差分プライバシーを満たさず，ポイズニング攻撃を受けたモデル，(4) 差分プライバシーを満たし，ポイズニング攻撃を受けたモデルへのメンバーシップ推定攻撃の成功率を比較する．

ポイズニング攻撃を踏み台にすることでメンバーシップ推定攻撃の成功率は上昇するため [11], [12]，(1) より (3) の攻撃成功率が高くなる．また，一般に差分プライバシーによってメンバーシップ推定攻撃の攻撃成功率は低下するため [44], [61]，(1) より (2) の攻撃成功率が低くなる．

これに対し，本稿ではまず差分プライバシーを適用したモデルについても同様に，ポイズニング攻撃によってメンバーシップ推定攻撃の成功率が上昇するか，つまり (2) より (4) の方が攻撃成功率が高くなるかどうか検証する．さらに，差分プライバシーによってポイズニング支援型メンバーシップ推定攻撃の成功率が低下するか，つまり (3) より (4) の方が攻撃成功率が低くなるかどうか検証する．

このとき，汚染データのサンプル数と，差分プライバシーの敏感度を様々に変えて検証することで，どの程度の攻撃に対して，どの程度の差分プライバシーが有効かその関係性を明らかにする．

4.1 実験設定

モデルアーキテクチャは ResNet18，データセットは CIFAR-10，機械学習フレームワークは PyTorch を利用した．ハイパーパラメータは表 1 の通りである．

4.1.1 ポイズニング攻撃

ポイズニング攻撃は，文献 [12] の手法を参考に実装した．不正なラベルを含むデータを，被害モデルの学習データに混入することで，被害モデル全体の学習を阻害する．

具体的には，CIFAR-10 の学習データセット 50,000 枚の中から，250 枚をランダムに抽出し，不正なラベルに入れ替えて汚染データとする．汚染データ 250 枚を $1, 2, 4, 8, 16$ セット数だけ用意し，$50,000 - 250 = 49,750$ 枚と組み合わせて被害モデルに学習させる．

表 1 学習時のハイパーパラメータ

	Learning Rate	Epoch	δ	C
Non DP	0.1 w/ scheduler	200	-	-
DP	1e-3	200	1e-5	1.0

学習時のパラメータは，表 1 に記載されている汚染データなしで学習した場合の設定にしたがう．

4.1.2 差分プライバシー

差分プライバシーは PyTorch の Opacus [49][*1]を利用する．これは画像にノイズを加えて学習することで，学習データのプライバシーを保護することができる．

3.2 節で述べた ϵ が異なるデータセット間の距離を定義し，δ で偶然にプライバシーが漏洩する確率を定義する．つまり，ϵ が小さければ小さいほど，異なるデータセット間の距離は小さくなり，より強固なプライバシーが保証される．また，ϵ が小さいほど敏感度も小さくなる．クリッピングパラメータ C の大きさでデータごとの勾配の大きさを制限する．今回は，文献 [58] を参考に，δ を固定して，ϵ を各エポックに対して $0.2, 1.0, 1.8$ に設定している．

CIFAR-10 データセットを用いて，表 1 の設定で学習を行った．これらの設定は，Opacus が提供するサンプルコードを参考にしている．

4.1.3 メンバーシップ推定攻撃

メンバーシップ推定攻撃は Shokri らの方式 [62] を実装した．コードは ART ライブラリ[*2]を利用している．

被害モデルの学習データセットである CIFAR-10 50,000 枚をランダムに 10 分割した． この 10 分割された各 5,000 枚と，被害モデルの学習には用いない CIFAR-10 のテストデータセット 10,000 枚をランダムに 2 分割した 5,000 枚を合わせて 10,000 枚とする．この 10,000 枚を用いて攻撃者の持つ攻撃モデルを学習する．

このとき攻撃モデルは，各データの正解ラベルと，被害モデルが推論したラベル，およびそのデータが被害モデルの学習に使われたかどうかの情報を用いて学習する．また，攻撃モデルの学習に用いず，被害モデルの学習に用いたデータ 5,000 枚と，攻撃モデルの学習にも被害モデルの学習にも用いない 5,000 枚で，攻撃モデルのテストを行う．

本実験では上述した設定の下，10 分割した CIFAR-10 の学習データセットと，2 分割したテストデータセットの組み合わせ全 20 種に対して交差検証を行った．

4.2 実験結果

CIFAR-10 で学習した ResNet18 に対し，差分プライバ

[*1] https://github.com/pytorch/opacus
[*2] https://github.com/Trusted-AI/adversarial-robustness-toolbox

表 2 モデルのテスト精度

		汚染データのセット数					
		0	1	2	4	8	16
ϵ	-	95.37	87.96	86.52	88.02	87.66	87.09
	1.8	51.38	57.47	56.90	58.61	57.61	57.91
	1.0	50.72	58.37	57.63	58.70	56.72	56.58
	0.2	51.11	56.08	57.05	54.75	55.85	55.77

表 3 メンバーシップ推定攻撃の攻撃成功率

		汚染データのセット数					
		0	1	2	4	8	16
ϵ	-	57.77	83.03	75.04	72.06	72.71	76.07
	1.8	55.42	54.79	55.56	54.64	55.27	55.25
	1.0	54.60	54.86	55.54	55.33	54.84	54.82
	0.2	54.38	54.13	54.68	54.46	54.80	54.60

シー，およびポイズニング攻撃を行った場合の，メンバーシップ推定攻撃の成功率を表 3 に示す．各モデルのテスト精度は表 2 の通りである．メンバーシップ推定攻撃は，攻撃モデルの学習に用いるデータを 20 種類用意し，交差検証を行った．攻撃成功率 Accuracy は，被害モデルの学習データを判別するタスクに対して以下の式で算出する．

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

ポイズニング攻撃によって攻撃成功率が上昇するか：差分プライバシーを満たさないモデルについては，ポイズニング支援型メンバーシップ推定攻撃の成功率が，支援型でない場合の攻撃成功率よりも高いことがわかる．

一方で，差分プライバシーを満たすモデルについては，ポイズニング支援型メンバーシップ推定攻撃の成功率が，必ずしもより高いわけではない．

差分プライバシーによって攻撃成功率が低下するか：差分プライバシーを満たすモデルへのポイズニング支援型メンバーシップ推定攻撃の成功率が，差分プライバシーを満たさないモデルへの攻撃成功率よりも低いことがわかる．

このとき，差分プライバシーを満たすモデルへの攻撃成功率は，ポイズニング支援型メンバーシップ推定攻撃の場合に比べて，支援型でない場合では最大 0.94% しか上昇していない．このことから，$\epsilon \leq 360$ の差分プライバシーであれば，ポイズニング支援型メンバーシップ推定攻撃の成功率の上昇を 0.94% に抑えることができる．

5. 考察

5.1 汚染データのセット数と攻撃成功率の影響

本実験では，差分プライバシーを適用したモデルではポイズニング攻撃による攻撃成功率の上昇が観測できなかった．ポイズニング支援型メンバーシップ推定攻撃では，ポイズニング攻撃が被害モデルに対して何かしらの変化を与えることで，メンバーシップ推定攻撃の攻撃成功率を上昇させていると考えられる．そのため，実験結果を考慮すると，ポイズニング攻撃が被害モデルに与える変化を差分プライバシーが抑制していると考えられる．

本実験で用いた被害モデルの学習ロスを図 1 および図 2 に示す．図 1 は差分プライバシーを適用しないモデルが汚染データを学習した際のロスであり，エポック数が 120 から 160 の地点でロスの値が跳ね上がっていることがわかる．このピークの後，急激にロスの値が低下していることから，無作為の汚染データの規則性を発見できずにいたモデルの学習が，この地点で一気に収束したと考えられる．

図 1 で示されたロスのピーク地点と，表 3 の攻撃成功率を対応させると，より少ないエポック数でロスがピークを迎えているモデルの方が，より高い攻撃成功率を示すことがわかる．これは，セット数が 2，8 の場合も同様であった．一方で，ピークの最大値と攻撃成功率の間に相関関係は見られなかった．

また，図 2 は差分プライバシーを適用したモデルが汚染データを学習した際のロスであり，図 1 で見られたようなロスのピークは観測できない．したがって，差分プライバシーを満たすことでポイズニング攻撃によるロスのピークの発生を抑制することができると考えられる．この結果は差分プライバシーによりポイズニング攻撃を抑制する既存の結果と一致している [52], [53].

既知の結果 [44], [61] として差分プライバシーがメンバーシップ推定攻撃を防ぐことと併せて考えると，差分プライバシーの適用はポイズニング支援型メンバーシップ推定攻撃の対策に有効と結論付ける．

5.2 対策の重ね掛け

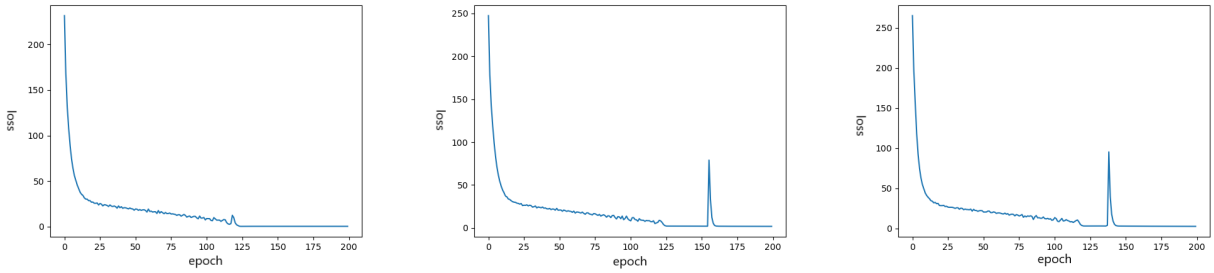
支援型攻撃への対策として単純に考えると，踏み台にされる攻撃，および本命の攻撃のどちらに対しても適切な対策を施せば防ぐことが可能である．2.1 節で述べたように，支援型攻撃として，モデル抽出攻撃，ポイズニング攻撃，敵対的サンプル，またはデータ復元攻撃を組み合わせる手法が提案されている．

これら全ての攻撃に対して，差分プライバシーなどの対策を施すことで支援型攻撃を防げるように思える．しかし，敵対的サンプルおよびデータ復元攻撃に対する攻撃対策はトレードオフの関係にあることがわかっている [63]. このように，複数の攻撃対策を重ねて施すことは困難であるため，支援型攻撃に適した攻撃対策が求められる．

5.3 潜在的な対策

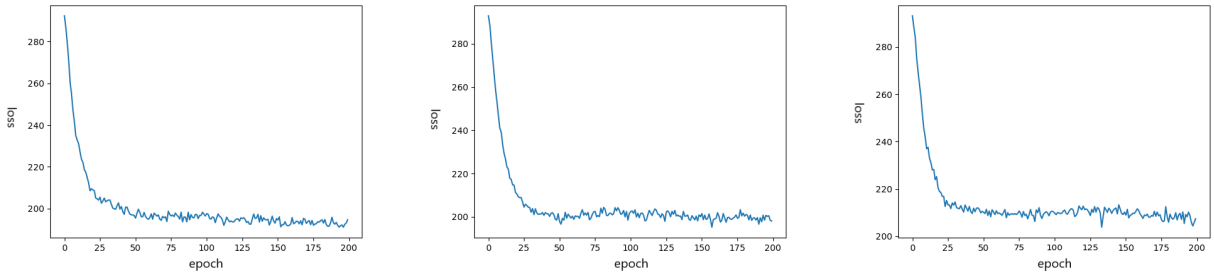
5.1 節で述べたように，被害モデルの学習におけるロスのピークの有無がメンバーシップ推定攻撃の攻撃成功率に影響があると考えられる．5.1 節では差分プライバシーが対策として有効なことを述べたが，以降ではそれ以外の可能性も検討する．

特筆すべき点として，差分プライバシーではロスのピー



(a) $PoisonSet = 1$　(b) $PoisonSet = 4$　(c) $PoisonSet = 16$

図 1　差分プライバシーなしでポイズニング攻撃を受けたモデルの学習ロス



(a) $PoisonSet = 1$　(b) $PoisonSet = 4$　(c) $PoisonSet = 16$

図 2　差分プライバシーありでポイズニング攻撃を受けたモデルの学習ロス

クの発生を抑制したが，勾配の急な上昇を検知する手法 [64], [65] を用いれば，ロスのピークの有無を把握できる．この手法を用いれば，メンバーシップ推定攻撃が成功しやすいモデルを判別することができ，メンバーシップ推定攻撃による被害を防ぐことができる可能性が高い．

5.4　制約条件

本稿では，1 種類のモデルアーキテクチャおよびデータセットを用いて実験を行った．また今回の実験では，ポイズニング攻撃として無差別型，データ復元攻撃としてシャドウモデルを用いないメンバーシップ推定を採用した．加えて，各モデルについて試行を繰り返していないことからも，汎用的な実験結果とは言えない．

5.5　今後の課題

被害モデルの学習におけるロスのピークの有無がメンバーシップ推定攻撃の攻撃成功率に影響すると考えられる．これを明らかにするためには，ロスのピーク前後の攻撃成功率を測定する必要がある．

また，より少ないエポック数でロスのピークが発生する条件を調査し，より強力なポイズニング支援型メンバーシップ推定攻撃に対して，差分プライバシーをはじめとする攻撃対策手法で防ぐことを考える．

6.　結論

本稿では，ポイズニング支援型メンバーシップ推定攻撃に対して，被害モデルが差分プライバシーを満たすことで攻撃を防ぐことが可能かどうか調査した．その結果，50,000 枚の学習データのうち 250 枚の汚染データによって，メンバーシップ推定攻撃の攻撃成功率が 25.26% 上昇することがわかった．一方で，$\epsilon \leq 360$ の差分プライバシーを満たすことで，メンバーシップ推定攻撃の攻撃成功率の上昇を 0.94% 以内に抑えることに成功した．

また，ポイズニング攻撃によって被害モデルの学習ロスが瞬間的に上昇する現象が観測された．一方で，差分プライバシーを満たす被害モデルについては，この瞬間的な上昇は観測されなかった．このとき，ポイズニング攻撃によってメンバーシップ推定攻撃の成功率は上昇していないことから，ロスの瞬間的な上昇がメンバーシップ推定攻撃の成功率に影響があると考えられる．

したがって，差分プライバシーを満たすモデルについては，ポイズニング支援型メンバーシップ推定攻撃を防ぐことができると結論付ける．

参考文献

[1] Biggio, B., Nelson, B. and Laskov, P.: Poisoning Attacks against Support Vector Machines, *Proc. of ICML 2012*, Omnipress, pp. 1467–1474 (2012).

[2] Charikar, M., Steinhardt, J. and Valiant, G.: Learning from Untrusted Data, *Proc. of STOC 2017*, ACM, pp. 47–60 (2017).

[3] Geiping, J., Fowl, L. H., Huang, W. R., Czaja, W., Taylor, G., Moeller, M. and Goldstein, T.: Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching, *Proc. of ICLR 2021*, (online), available from

⟨https://openreview.net/forum?id=01olnfLIbD⟩ (2021).

[4] Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T. and Goldstein, T.: Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks, *Proc. of NeurIPS 2018*, Vol. 31, Curran Associates, Inc., pp. 6106–6116 (2018).

[5] Gu, T., Liu, K., Dolan-Gavitt, B. and Garg, S.: BadNets: Evaluating Backdooring Attacks on Deep Neural Networks, *IEEE Access*, Vol. 7, pp. 47230–47244 (2019).

[6] Shokri, R., Stronati, M., Song, C. and Shmatikov, V.: Membership Inference Attacks Against Machine Learning Models, *Proc. of IEEE S&P 2018*, IEEE Computer Society, pp. 3–18 (2017).

[7] Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M. and Backes, M.: ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models, *Proc. of NDSS 2019*, The Internet Society (2019).

[8] Carlini, N., Liu, C., Erlingsson, Ú., Kos, J. and Song, D.: The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks, *Proc. of USENIX Security*, USENIX Association, pp. 267–284 (2019).

[9] Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A. and Raffel, C.: Extracting Training Data from Large Language Models, *Proc. of USENIX Security 2021*, USENIX Association, pp. 2633–2650 (2021).

[10] Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D. and Ristenpart, T.: Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing, *Proc. of USENIX Security 2014*, USENIX Association, pp. 17–32 (2014).

[11] Mahloujifar, S., Ghosh, E. and Chase, M.: Property Inference from Poisoning, *Proc. of IEEE S&P 2022*, IEEE, pp. 1569–1569 (2022).

[12] Tramèr, F., Shokri, R., Joaquin, A. S., Le, H., Jagielski, M., Hong, S. and Carlini, N.: Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets, *arXiv preprint arXiv:2204.00032* (2022).

[13] HIDANO, S., MURAKAMI, T., KATSUMATA, S., KIYOMOTO, S. and HANAOKA, G.: Model Inversion Attacks for Online Prediction Systems: Without Knowledge of Non-Sensitive Attributes, *IEICE Transactions on Information and Systems*, Vol. E101.D, No. 11, pp. 2665–2676 (2018).

[14] Dwork, C.: Differential Privacy, *Proc. of ICALP*, LNCS, Vol. 4052, Springer, pp. 1–12 (2006).

[15] Tramér, F., Zhang, F. and Juels, A.: Stealing Machine Learning Models via Prediction APIs, *Proc. of USENIX Security 2016*, USENIX Association, pp. 601–618 (2016).

[16] Juuti, M., Szyller, S., Marchal, S. and Asokan, N.: PRADA: Protecting against DNN Model Stealing Attacks, *Proc. of EuroS&P 2019*, IEEE, pp. 512–527 (2019).

[17] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Berkay Celik, Z. and Swami, A.: Practical Black-Box Attacks Against Machine Learning, *Proc. of AsiaCCS 2017*, ACM, pp. 506–519 (2017).

[18] He, X., Lyu, L., Sun, L. and Xu, Q.: Model Extraction and Adversarial Transferability, Your BERT is Vulnerable!, *Proc. of NAACL-HLT 2021*, ACL, pp. 2006–2012 (2021).

[19] Yue, Z., He, Z., Zeng, H. and McAuley, J. J.: Black-Box Attacks on Sequential Recommenders via Data-Free Model Extraction, *Proc. of RecSys 2021*, ACM, pp. 44–54 (2021).

[20] Huster, T. and Ekwedike, E.: TOP: Backdoor detection in neural networks via transferability of perturbation, *arXiv preprint arXiv:2103.10274* (2021).

[21] Lyu, L., He, X., Wu, F. and Sun, L.: Killing Two Birds with One Stone: Stealing Model and Inferring Attribute from BERT-based APIs, *CoRR*, Vol. abs/2105.10909 (online), available from ⟨https://arxiv.org/abs/2105.10909⟩ (2021).

[22] Gong, X., Chen, Y., Yang, W., Mei, G. and Wang, Q.: InverseNet: Augmenting Model Extraction Attacks with Training Data Inversion., *IJCAI*, pp. 2439–2447 (2021).

[23] Fowl, L., Goldblum, M., Chiang, P.-y., Geiping, J., Czaja, W. and Goldstein, T.: Adversarial Examples Make Strong Poisons, *Proc. of NeurIPS 2021*, Vol. 34, Curran Associates, Inc., pp. 30339–30351 (2021).

[24] Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C. and Li, B.: Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning, *Proc. of IEEE S&P 2018*, IEEE, pp. 19–35 (2018).

[25] Muñoz González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E. C. and Roli, F.: Towards Poisoning of Deep Learning Algorithms with Back-Gradient Optimization, *Proc. of AISec 2017*, ACM, pp. 27–38 (2017).

[26] Suciu, O., Marginean, R., Kaya, Y., III, H. D. and Dumitras, T.: When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks, *Proc. of USENIX Security 2018*, USENIX Association, pp. 1299–1316 (2018).

[27] Liu, Y., Ma, S., Aafer, Y., Lee, W., Zhai, J., Wang, W. and Zhang, X.: Trojaning Attack on Neural Networks, *Proc. of NDSS 2018*, The Internet Society, (online), available from ⟨http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/ndss2018_03A-5_Liu_paper.pdf⟩ (2018).

[28] Tan, T. J. L. and Shokri, R.: Bypassing Backdoor Detection Algorithms in Deep Learning, *Proc. of EuroS&P*, IEEE, pp. 175–183 (2020).

[29] Wang, B., Yao, Y., Shan, S., Li, H., Viswanath, B., Zheng, H. and Zhao, B. Y.: Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks, *IEEE S&P 2019*, IEEE, pp. 707–723 (2019).

[30] Chou, E., Tramèr, F. and Pellegrino, G.: SentiNet: Detecting Localized Universal Attacks Against Deep Learning Systems, *Proc. of IEEE SPW 2020*, IEEE, pp. 48–54 (2020).

[31] Tran, B., Li, J. and Madry, A.: Spectral Signatures in Backdoor Attacks, *Proc. of NIPS 2018*, ACM, p. 8011–8021 (2018).

[32] Chen, B., Carvalho, W., Baracaldo, N., Ludwig, H., Edwards, B., Lee, T., Molloy, I. M. and Srivastava, B.: Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering, *Proc. of SafeAI 2019* (2019).

[33] Li, S., Xue, M., Zhao, B., Zhu, H. and Zhang, X.: Invisible Backdoor Attacks on Deep Neural Networks via Steganography and Regularization, *IEEE Transactions on Dependable and Secure Computing*, Vol. 18, No. 05, pp. 2088–2105 (2021).

[34] Ning, R., Li, J., Xin, C. and Wu, H.: Invisible Poison: A Blackbox Clean Label Backdoor Attack to Deep Neural Networks, *Proc. of INFOCOM 2021*, IEEE (2021).

[35] Li, S., Liu, H., Dong, T., Zhao, B. Z. H., Xue, M., Zhu, H. and Lu, J.: Hidden Backdoors in Human-Centric Language Models, *Proc. of CCS 2021*, ACM, pp. 3123–3140 (2021).

[36] Zhong, H., Liao, C., Squicciarini, A. C., Zhu, S. and Miller, D.: Backdoor Embedding in Convolutional Neural Network Models via Invisible Perturbation, *Proc. of CODASPY 2020*, ACM, pp. 97–108 (2020).

[37] Doan, K., Lao, Y. and Li, P.: Backdoor Attack with Imperceptible Input and Latent Modification, *Proc. of NeurIPS 2021*, Vol. 34, Curran Associates, Inc., pp. 18944–18957 (2021).

[38] Bagdasaryan, E. and Shmatikov, V.: Blind Backdoors in Deep Learning Models, *Proc. of USENIX Security 2021*, USENIX Association (2021).

[39] Song, C., Ristenpart, T. and Shmatikov, V.: Machine Learning Models That Remember Too Much, *Proc. of CCS 2017*, ACM, pp. 587–601 (2017).

[40] Boenisch, F., Dziedzic, A., Schuster, R., Shamsabadi, A. S., Shumailov, I. and Papernot, N.: When the Curious Abandon Honesty: Federated Learning Is Not Private, *CoRR*, Vol. abs/2112.02918 (online), available from ⟨https://arxiv.org/abs/2112.02918⟩ (2021).

[41] Fowl, L., Geiping, J., Reich, S., Wen, Y., Czaja, W., Goldblum, M. and Goldstein, T.: Decepticons: Corrupted Transformers Breach Privacy in Federated Learning for Language Models, *CoRR*, Vol. abs/2201.12675 (online), available from ⟨https://arxiv.org/abs/2201.12675⟩ (2022).

[42] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A. and Tramèr, F.: Membership Inference Attacks From First Principles, *Proc. of IEEE S&P 2022*, IEEE, pp. 1897–1914 (2022).

[43] Fredrikson, M., Jha, S. and Ristenpart, T.: Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures, *Proc. of CCS 2015*, ACM, pp. 1322–1333 (2015).

[44] Yeom, S., Giacomelli, I., Fredrikson, M. and Jha, S.: Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting, *Proc. of CSF 2018*, IEEE, pp. 268–282 (2018).

[45] Henderson, P., Sinha, K., Angelard-Gontier, N., Ke, N. R., Fried, G., Lowe, R. and Pineau, J.: Ethical Challenges in Data-Driven Dialogue Systems, *Proc. of AIES 2018*, ACM, pp. 123–129 (2018).

[46] Melis, L., Song, C., De Cristofaro, E. and Shmatikov, V.: Exploiting Unintended Feature Leakage in Collaborative Learning, *IEEE S&P 2019*, IEEE, pp. 691–706 (2019).

[47] Nasr, M., Shokri, R. and Houmansadr, A.: Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning, *IEEE S&P 2019*, IEEE, pp. 739–753 (2019).

[48] Wen, Y., Geiping, J. A., Fowl, L., Goldblum, M. and Goldstein, T.: Fishing for User Data in Large-Batch Federated Learning via Gradient Magnification, *Proc. of ICML 2022*, PMLR, Vol. 162, PMLR, pp. 23668–23684 (2022).

[49] Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Gosh, S., Bharadwaj, A., Zhao, J., Cormode, G. and Mironov, I.: Opacus: User-Friendly Differential Privacy Library in PyTorch, *CoRR*, Vol. abs/2109.12298 (online), available from ⟨https://arxiv.org/abs/2109.12298⟩ (2021).

[50] Papernot, N.: Machine Learning at Scale with Differential Privacy in TensorFlow, *Proc. of PEPR 2019*, USENIX Association, (online), available from ⟨https://www.usenix.org/node/238163⟩ (2019).

[51] Prediger, L., Loppi, N. A., Kaski, S. and Honkela, A.: d3p - A Python Package for Differentially-Private Probabilistic Programming, *Proceedings on Privacy Enhancing Technologies*, Vol. 2022, No. 2, pp. 407–425 (2022).

[52] Ma, Y., Zhu, X. and Hsu, J.: Data Poisoning against Differentially-Private Learners: Attacks and Defenses, *Proc. of IJCAI 2019*, ijcai.org, pp. 4732–4738 (2019).

[53] Du, M., Jia, R. and Song, D.: Robust anomaly detection and backdoor attack detection via differential privacy, *Proc. of ICLR 2020*, (online), available from ⟨https://openreview.net/forum?id=SJx0q1rtvS⟩ (2020).

[54] Doan, B. G., Abbasnejad, E. and Ranasinghe, D. C.: Februus: Input Purification Defense Against Trojan Attacks on Deep Neural Network Systems, *Proc. of ACSAC 2020*, ACM, p. 897–912 (2020).

[55] Veldanda, A. K., Liu, K., Tan, B., Krishnamurthy, P., Khorrami, F., Karri, R., Dolan-Gavitt, B. and Garg, S.: NNoculation: Catching BadNets in the Wild, *Proc. of AISec 2021*, ACM, p. 49–60 (2021).

[56] Cao, X., Jia, J. and Gong, N. Z.: Data Poisoning Attacks to Local Differential Privacy Protocols, *Proc. of USENIX Security 2021*, USENIX Association, pp. 947–964 (2021).

[57] Cheu, A., Smith, A. and Ullman, J.: Manipulation Attacks in Local Differential Privacy, *Proc. of IEEE S&P 2021*, IEEE, pp. 883–900 (2021).

[58] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K. and Zhang, L.: Deep Learning with Differential Privacy, *Proc. of CCS 2016*, ACM, pp. 308–318 (2016).

[59] Jayaraman, B. and Evans, D.: Evaluating Differentially Private Machine Learning in Practice, *Proc. of USENIX Security 2019*, USENIX Association, pp. 1895–1912 (2019).

[60] Jagielski, M., Ullman, J. and Oprea, A.: Auditing Differentially Private Machine Learning: How Private is Private SGD?, *Advances in Neural Information Processing Systems*, Vol. 33, Curran Associates, Inc., pp. 22205–22216 (2020).

[61] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K. and Zhang, L.: Deep learning with differential privacy, *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318 (2016).

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

[63] Song, L., Shokri, R. and Mittal, P.: Privacy risks of securing machine learning models against adversarial examples, *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 241–257 (2019).

[64] Chan, A. and Ong, Y.-S.: Poison as a cure: Detecting & neutralizing variable-sized backdoor attacks in deep neural networks, *arXiv preprint arXiv:1911.08040* (2019).

[65] Gao, Y., Doan, B. G., Zhang, Z., Ma, S., Zhang, J., Fu, A., Nepal, S. and Kim, H.: Backdoor attacks and countermeasures on deep learning: A comprehensive review, *arXiv preprint arXiv:2007.10760* (2020).