An Equivalence Between Private Classification and Online Prediction

Mark Bun [1] , Roi Livni [2], Shay Moran [3]

[1]Boston University [2]Tel Aviv University [3]Technion

Presenter: Yucheng Sun, Jiaqing Xie

ETH zürich

Table of Contents

Main Question

Which tasks can be learned while protecting private data

Answer: Any task which is online learnable

differential privacy: young area
online learning: mature area

Online & Private Learning

Online Learning and Private Learning are closely related.

"Differential privacy may be ensured for free in online linear optimization" [AS17]

"DP-inspired stability is well-suited to designing online learning algorithms with excellent guarantees" [Abe+19]

"Differential privacy is enabled by stability and ensures stability" [DR+14]

Proof of Equivalence

Two directions:

1) Private Learnable $\Rightarrow$ Online Learnable

2) Online Learnable$\Rightarrow$ Private Learnable

1) Private Learnable ⇒ Online Learnable

One direction: Private Learnable ⇒ Finite Littlestone dimension ⇒ Online Learnable

1) Private Learnable ⇒ Finite Littlestone dimension ✓

"Every approximately differentially private learning algorithm for a class H with Littlestone dimension d requires $\Omega(\log^*(d))$ examples" [Alo+19]

2) Finite Littlestone dimension ⇒ Online Learnable ✓

"Hypothesis is online learnable iff it has a finite Littlestone dimension" [BPS09]

2) Online Learnable $\Rightarrow$ Private Learnable ?

Another direction: Online Learnable $\Rightarrow$ Finite LittleStone dimension $\Rightarrow$ Private Learnable

1) Online Learnable $\Rightarrow$ Finite LittleStone dimension ✓
"Hypothesis is online learnable iff it has a finite Littlestone dimension"

2) Finite LittleStone dimension $\Rightarrow$ Private Learnable: If true, then the equivalence is proved.

Idea: introduce stability. We prove 1) Finite LittleStone dimension $\Rightarrow$ Stability $\Rightarrow$ Private Learnable [This paper]

Overview

- Online learning, Littlestone dimension and SOA
- Stability
- Finite Littlestone dimension $\Rightarrow$ Stability
- Differential Privacy
- Stability $\Rightarrow$ Private learnable

Table of Contents

Online Learning

Definition (Online Learning)

Hypothesis class $H = \{h : X \to \{\pm 1\}\}$. Examples of inputs $(x_1, y_1), \ldots,$ $(x_n, y_n) \in X \times \{\pm 1\}$.

For each time step t

1. observe one instance x_t from the examples
2. predict the label $\hat{y}_t$ with $h \in H$
3. receive the true label y_t
4. compute the $loss(\hat{y}_t, y_t)$
5. update H with some metrics
6. continue till t = n

Online Learning

Definition (Goal of Online Learning)

We want to minimize the regret, the number of mistakes compared to the best hypothesis in H:

$$R(n) = \sum_{t=1}^{n} 1[y_t \neq \hat{y}_t] - \min_{h^* \in H} \sum_{t=1}^{n} 1[y_t \neq h^*]$$

Theorem (Boundedness of R(n))

R(n) has been proved to be bounded:

$$\Omega(\sqrt{dn}) \leq R(n) \leq O(\sqrt{dn \log n})$$

where d is the Littlestone Dimension

Online Learnable $\Leftrightarrow$ Finite Littlestone dimension

1. Boundedness of R(n) $\Leftrightarrow$ hypothesis class H is online learnable with finite littlestone dimension

2. Next step: littlestone dimension and mistake bound

Littlestone Dimension

Definition (Mistake bound)

Number of possible mistakes that an online learning algorithm A made during prediction. Let's denote it by $M_A(H)$

Theorem (Upper-Boundedness)

Consistent algorithm returns a $|H| - 1$ *upper bound. Halving algorithm returns a* $\log_2 |H|$ *upper bound.*

Definition (Lower-Boundedness $\Rightarrow$ LittleStone Dimension)

The best achievable lower bound for $M_A(H)$, which is denoted by LDim(H). For any online learning algorithm A, we have:

$$M_A(H) \geq LDim(H)$$

Littlestone Dimension

Intuition:

1) We view online-learning as the game between the gamer and the environment. The environment wants the gamer to make mistakes. Under the settings of a 0-1 binary classification scenario, if a learner picks the prediction y_t, the environment will pick $1 - y_t$

2) How to make maximum mistakes? Answer: Build a complete binary tree where each sample corresponds to one node in the tree. If a prediction given by gamer is 0, then environment would give 1 (right prediction) and travel to left child

3) Example:

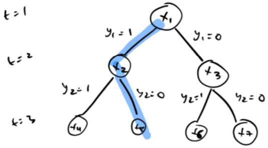

Figure: Complete binary tree where LDim = 2 (tree depth T)

Standard Optimal Algorithm (SOA)

Algorithm (Standard Optimal Algorithm)

The best achievable mistake lower bound is achievable by implementing Standard Optimal Algorithm (SOA), where we have:

$$M_A(H) = LDim(H)$$

For each time step t

1. *observe one instance x_t from the examples*
2. *choose $b \in \{\pm 1\}$, let $H' = \{h \in H : h(x_t) = b\}$. We predict $h' = \arg\max_b LDim(H')$*
3. *receive the true label y_t*
4. *compute the loss*
5. *Update H*

Standard Optimal Algorithm (SOA)

Definition (Realizable to non-realizable samples)

So far samples could be shattered by H $\Leftrightarrow$ Samples are realizable.

1. If the incoming sample x_{t+1}, y_{t+1} maintains the realizability, update H by SOA.
2. Else:

$$h'(x_{t+1}) = y_{t+1}$$

while keeping other $h'(\cdot)$ unchanged.

Milestone 1: Online Learner

Current:

- Hypothesis with finite Littlestone dimension is online-learnable
- Optimal online-learnable learning algorithm w.r.t. mistake bound model is given by **SOA**

Next step:

- **SOA** is used in stable learner as an algorithm
- introduce stability

Stability

Definition (Global Stability)

Let n be the sample size and η a stability parameter. Algorithm A is (n, η) globally stable w.r.t D if there exists a hypothesis h s.t.

$$\mathbb{P}_{S \sim D^n}[A(S) = h] \geq \eta$$

Stability $\Rightarrow$ Generalization

Proposition (Global Stable Learner $\Rightarrow$ generalize well)

Suppose A is realizable learner ($loss(A(S)) = 0$ for any realizable S). D is a realizable distribution s.t. A is (n, η) stable and $\underset{S \sim D^n}{\mathbb{P}}[A(S) = h] \geq \eta$ for a hypothesis h, then

$$loss_D(h) \leq \frac{\ln(1/\eta)}{n}$$

Milestone 2: Online + Stable Learner

- A hypothesis class H has a finite LDim d,
- A = SOA , h $\in$ H, $\underset{S\sim D^n}{\mathbb{P}}[SOA(S) = h] \geq \eta$, $loss_D(h) \leq \frac{\ln(1/\eta)}{n}$

Intuition:

Proposition (Finite LDim)

Suppose the hypothesis class has a finite LDim d. If we could find specific finite η and n, using SOA, properties of stability and generalization of a stable learner are ensured.

The direction Finite Littlestone $\Rightarrow$ Global stable learner ✓

Finite Littlestone $\Rightarrow$ Global stable learner

Theorem

Let H be a hypothesis class with Littlestone dimension $d \leq 1$, let $\alpha > 0$, set number of samples

$$n = 2^{2^{d+2}+1} 4^{d+1} \lceil \frac{2^{d+2}}{\alpha} \rceil$$

D is a realizable distribution and $S \sim D^n$, then there exists a randomized algorithm G: $X \times \{\pm 1\} \rightarrow \{\pm 1\}^X$ and a hypothesis $f \in H$: with the following properties:

$$\mathbb{P}[G(S) = f] \geq \frac{1}{(d+1)2^{2^d+1}} \quad \text{and} \quad loss_D(f) \leq \alpha$$

Distribution $\mathcal{D}_k$

Q: How are samples randomly chosen from original distribution D ?
A: The sampling depends on **tournament samples**
Aim: To make sure that each round the algorithm will make a mistake

Algorithm (Choosing D_k)

$D_k = D_k(k, n)$ are defined by induction on k:
D_0: outputs the empty sample $\emptyset$ with prob. 1 For each k:

1. *Draw $S_0, S_1 \sim D_{k-1}$ and $T_0, T_1 \sim D^n$*
2. *$f_0 = SOA(S_0 \circ T_0)$, $f_1 = SOA(S_1 \circ T_1)$, $\circ$ means append*
3. *If $f_0 = f_1$, GOTO 1*
4. *Pick $x \in \{x : f_0(x) \neq f_1(x)\}$ and sample $y \sim \{\pm 1\}$ uniformly*
5. *If $f_0(x) \neq y$, output $S_0 \circ T_0 \circ \mathbf{(x,\ y)}$, else $S_1 \circ T_1 \circ \mathbf{(x,\ y)}$*

k mistakes were made. $SOA(S \circ T)$ is consistent with T (realizable).

Monte-Carlo Variants of Distribution $\mathcal{D}_k$

Problem : From 1) to 3), it might generate infinite number of samples (unbounded) from D.

Solution : In order to circumvent being stuck in the sampling period, if more than N examples are sampled from D, break immediately and output Fail.

Milestone 3: Prove Finite littlestone dimension $\Rightarrow$ Stability

Finite littlestone dimension $\Rightarrow$ Stable learner

- prove stability
- prove stability $\Rightarrow$ generalization
- prove number of generated samples is finite
- Optimal f^* in hypothesis that looses the lower bound

Existence of Frequent Hypothesis

Lemma (global stability)

There exists $k \leq$ LDim d and an hypothesis $f : X \to \{\pm 1\}$ s.t.

$$\mathop{\mathbb{P}}_{S \sim D_k, T \sim D^n}[SOA(S \circ T) = f] \geq \eta = 2^{-2^{d+2}}$$

Proof: Existence of Frequent Hypothesis (Skip)

Sketch Proof (by contradiction):

1) Suppose $\underset{S\sim D_d, T\sim D^n}{\mathbb{P}}[SOA(S\circ T) = f] < 2^{-2^{d+2}}$

2) ρ_k: prob. that all k tournament examples are consistent with c where c is the target concept.

3) two cases: $S_0, S_1 \in D_{k-1}$ are all consistent with c, therefore, each ρ_{k-1} leads to ρ_{k-1}^2. And $f_0 \neq f_1$. We have that $\mathbb{P}[f_0 = f_1] < 2^{-2^{d+2}} < 8 \cdot 2^{-2^{d+2}}$ according to 1).

4) conditioned on uniformly distributed y with prob $1/2$. put it all together it will produce:

$$\rho_k \geq \frac{1}{2}(\rho_{k-1}^2 - 8 \cdot 2^{-2^{d+2}})$$

Proof: Existence of Frequent Hypothesis (Skip)

(Continue)

5) $\rho_0 = 1$, we can prove by induction that for $k \leq d, \rho_k \geq 4 \cdot 2^{-2^{k+1}}$
Suppose it stands for k-1, then for k

$$\begin{aligned}\rho_k &\geq \frac{\rho_{k-1}^2 - 8 \cdot 2^{-2^{d+2}}}{2} \\ &\geq \frac{(4 \cdot 2^{-2^k})^2 - 8 \cdot 2^{-2^{d+2}}}{2} \\ &= 8 \cdot 2^{-2^{k+1}} - 4 \cdot 2^{-2^{d+2}} \\ &\geq 4 \cdot 2^{-2^{k+1}} \quad \text{since k} \leq \text{d}\end{aligned}$$

6) so $\rho_k \geq 2^{-2^{d+2}}$. SOA does not make more than d mistakes, so if all tournament examples are consistent with c then $SOA(S \circ T) = f$, then $\underset{S\sim D_d, T\sim D^n}{\mathbb{P}}[SOA(S \circ T) = f] \geq 2^{-2^{d+2}}$, contradicts 1)

Proof

Finite littlestone dimension $\Rightarrow$ Stable learner

- **prove stability** $\checkmark$
- prove stability $\Rightarrow$ generalization
- prove number of generated samples is finite
- Optimal f^* in hypothesis that looses the lower bound

Generalization

Lemma (Generalization)

Let k be such that D_k is defined. Then every f s.t.

$$\mathbb{P}_{S \sim D_k, T \sim D^n}[SOA(S \circ T) = f] \geq 2^{-2^{d+2}}$$

satisfies

$$loss_D(f) \leq \frac{2^{d+2}}{n}$$

Proof: Generalization (Skip)

Sketch Proof (similar to generalization proof before):
1) A be the event that $SOA(S \circ T) = f$, by definition $\mathbb{P}[A] \geq 2^{-2^{d+2}}$
2) B be the event that f is consistent with T, by definition $\mathbb{P}[B] = (1-\alpha)^n$ again if we let $\alpha = loss_D(f)$
3) We've mentioned when $SOA(S \circ T) = f$,then f must be consistent with T. So that $A \subseteq B$
4) solve the inequality: $2^{-2^{d+2}} \leq (1-\alpha)^n \leq e^{-\alpha n}$ leads to the generalization bound.

Proof

Finite littlestone dimension $\Rightarrow$ Stable learner

- **prove stability** ✓
- **prove stability** $\Rightarrow$ **generalization** ✓
- prove number of samples is finite
- Optimal f^* in hypothesis that looses the lower bound

Expected Sample Complexity

Lemma

Let k be s.t. D_k is well-defined. M_k denotes the number of examples of D drawn in the process of generating $S \sim D_k$. Then:

$$\mathbb{E}[M_k] \leq 4^{k+1} \cdot n$$

, where n is the number of inputs.

Proof skipped here.

Algorithm G

Algorithm

1. *Consider the distribution* $\hat{D}_k$*, where* $n = \lceil \frac{2^{d+2}}{\alpha} \rceil$ *(From generalization bound)*
2. *Sampling upper bound is set to* $2^{2^{d+2}+1} 4^{d+1} \cdot n$*. Inverse of stability bound times* $\mathbb{E}[M_k]$
3. *Draw* $k \in \{0, 1, ..., d\}$ *uniformly at random*
4. *Output* $h = SOA(S \circ T)$, $S \sim \hat{D}_k$, $T \sim D_n$

Proof

Finite littlestone dimension $\Rightarrow$ Stable learner

- **prove stability** ✓
- **prove stability** $\Rightarrow$ **generalization** ✓
- **prove number of generated samples is finite** ✓
- Optimal f^* in hypothesis that looses the lower bound

Put all things together

We leave : there exists a f such that :

$$\underset{S\sim\hat{D}_k, T\sim D^n}{\mathbb{P}}[SOA(S \circ T) = f] \geq \frac{2^{-2^{d+2}}}{d+1}$$

and

$$loss_D(f) \leq \alpha$$

We already achieved the sample size $n = 2^{2^{d+2}+1}4^{d+1} \cdot \lceil \frac{2^{d+2}}{\alpha} \rceil$.

Proof

- Assume optimal k^* and $f = f^*$, we have $loss_D(f^*) \leq loss_D(f) = \alpha$
- Markov's inequality:

$$\mathbb{P}[M_{k^*} > 2^{2^{d+2}+1} 4^{d+1} \cdot \lceil \frac{2^{d+2}}{\alpha} \rceil] \leq 2^{-2^{d+2}-1}$$

- $\underset{S \sim \hat{D}_{k^*}, T \sim D^n}{\mathbb{P}}[SOA(S \circ T) = f^*] = \underset{S \sim D_{k^*}, T \sim D^n}{\mathbb{P}}[SOA(S \circ T) = f^* \quad \text{and} \quad M_{k^*} < 2^{2^{d+2}+1} 4^{d+1} \cdot \lceil \frac{2^{d+2}}{\alpha} \rceil] \geq 2^{-2^d-1} \geq 2^{-2^{d+2}-1}$
- $k = k^*$ with probability $\frac{1}{d+1}$. Finish the proof.

Proof

Finite littlestone dimension $\Rightarrow$ Stable learner $\checkmark$

- **prove stability $\checkmark$**
- **prove stability $\Rightarrow$ generalization $\checkmark$**
- **prove number of generated samples is finite $\checkmark$**
- **Optimal f^* in hypothesis that looses the lower bound $\checkmark$**

Table of Contents

Differential Privacy (General definition)

Definition

A randomized algorithm $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{R}^k$ is said to be (ϵ, δ)-*differentially private* if for all measurable $S \subseteq \mathcal{R}^k$ and all neighboring datasets $x, y \in \mathcal{X}$:

$$\Pr(\mathcal{M}(x) \in S) \leq e^{\epsilon} \Pr(\mathcal{M}(y) \in S) + \delta\,,$$

where two datasets x, y are said to be *neighboring* if they only differ in one entry.

I.e., an attacker cannot confidently infer whether a sample is in the input dataset given the output of the algorithm.

When (ϵ, δ) decrease, the level of privacy increases.

Two nice property of differential privacy

Theorem (Post-processing)

Let $\mathcal{M}\colon \mathcal{X}^n \to \mathcal{R}^k$ be an (ϵ, δ)-differentially private algorithm. Let $f\colon \mathcal{R}^k \to \mathcal{R}^k$ be an arbitrary mapping. Then $f \circ \mathcal{M}$ is also (ϵ, δ)-differentially private.

Theorem (Composition theorem)

Let $\mathcal{M}_1\colon \mathcal{X}^n \to \mathcal{R}^k$ be an (ϵ_1, δ_1)-differentially private algorithm and $\mathcal{M}_2\colon \mathcal{X}^n \to \mathcal{R}^k$ be an (ϵ_2, δ_2)-differentially private-algorithm, then their combination $M_{1,2}(x) = (\mathcal{M}_1(x), \mathcal{M}_2(x))$ is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-differentially private.

Global stable learner $\Rightarrow$Private Learning

To prove this statement, the paper constructs a private learner

Constructed using existing algorithms as black boxes

Building block 1: **Generic Private Learner** [Kas+08]
Building block 2: **Stable Histograms** [Kor+09; BNS19]

Generic Private Learner

Theorem

Let $H \subseteq \{\pm 1\}^X$ *be a collection of hypothesis. For*

$$n = O\left(\frac{\log(|H|) + \log(1/\beta)}{\alpha\epsilon}\right)$$

there exists an $(\epsilon, \delta)-$*differentially private algorithm* $GenericLearner : (X \times \{\pm 1\})^n \to H$ *such that the following holds. Let* D *be a distribution over* $(X \times \{\pm 1\})$ *such that there exists* $h^* \in H$ *with* $loss_D(h^*) \leq \alpha$*. Then on input* $S \sim D^n$*, algorithm GenericLearner outputs, with probability at least* $1 - \beta$*, a hypothesis* $\hat{h} \in H$ *such that* $loss_D(\hat{h}) \leq 2\alpha$*.*

Stable Histogram

Theorem (Stable Histogram)

Let X be any data domain. For

$$n \geq O\left(\frac{\log(1/\eta\beta\delta)}{\eta\epsilon}\right)$$

there exists an (ϵ, δ)-differentially private algorithm Hist which, with probability at least $1-\beta$, on input $S = (x_1, ..., x_n)$ outputs a list $L \subset X$ and a sequence of estimates $\alpha \in [0,1]^{|L|}$ such that

- *Every x with $freqS(x) \geq \eta$ appears in L and*
- *For every $x \in L$, the estimate a_x satisfies $|a_x - freqS(x)| \leq \eta$*

Construction of a private learner

Algorithm

Require: Stable learner G, Stable Histogram algorithm Hist, generic learner GenericLearner

Step 1. Let $S_1, ..., S_k$ each consist of m i.i.d. samples from D. Run G on each batch of samples producing $h_1 = G(S_1), ..., h_k = G(S_k)$.

Step 2. Run the Stable Histogram algorithm Hist on input $H = (h_1, ..., h_k)$ using privacy parameters $(\epsilon/2, \delta)$ and accuracy parameters $(\eta/8, \beta/3)$, producing a list L of frequent hypotheses.

Step 3. Let S' consist of n' i.i.d. samples from D. Run GenericLearner(S') using the collection of hypotheses L with privacy parameter $(\epsilon/2, 0)$ and accuracy parameters $(\alpha/2, \beta/3)$ to output a hypothesis $\hat{h}$.

Proof of correctness

To prove the correctness of the algorithm, one need to prove:

1. The algorithm is (ϵ, δ) private.
2. The algorithm is accurate.

Proof of correctness (privacy part)

Proof.

We know Step 2 is $(\epsilon/2, \delta)-$differentially private, Step 3 is $(\epsilon/2, 0)-$differentially private.
By using composition theorem, we can immediately prove the algorithm is $(\epsilon/2 + \epsilon/2, \delta)-$differentially private. □

Proof of correctness (accuracy part)

Proof.

Assume the optimal classifier is h^* and $loss_D(h^*) = \alpha/2$.
1. Using standard generalization argument, one can show that w.p. at least $1 - \beta/3$,

$$|freq_H(h) - \Pr_{S \sim D^m}[G(s) = h]| \leq \frac{\eta}{8}$$

for every $h \in \{\pm 1\}^X$ as long as $k \geq O(\log(1/\beta)/\eta)$.
2. Conditioning on 1. By the correctness of the Stable Histogram algorithm, we know w.p. $1 - \beta/3$ HIST produces a list L containing h^* with the estimate $a_{h^*} \geq \eta - \eta/8 - \eta/8 = \frac{3}{4}\eta$.
3. By the correctness of private generic learner, *GenericLearner* can successfully find $\hat{h} \in L$ and $loss_D(\hat{h}) \leq \alpha$.

By using the union bound, we know the algorithm successfully find a correct solution w.p. $1 - \beta$. □

Global stable learner $\Rightarrow$Private Learning (formal statement)

Theorem

Let H be a concept class over data domain X. Let $G : (X \times \{\pm 1\})^m \to \{\pm 1\}^X$ be a randomized algorithm such that, for D a realizable distribution and $S \sim D^m$, there exists a hypothesis h such that $Pr[G(S) = h] \geq \eta$ and $loss_D(h) \leq \alpha/2$. Then for some

$$n = O\left(\frac{m \log(1/\eta\beta\delta)}{n\epsilon} + \frac{\log(1/\eta\beta)}{\alpha\epsilon}\right)$$

there exists an (ϵ, δ)-differentially private algorithm $M : (X \times \{\pm 1\})^m \to \{\pm 1\}^X$ which, given n i.i.d. samples from D, produces a hypothesis $h\hat{}$ such that $loss_D(\hat{h}) \leq \alpha$ with probability at least $1 - \beta$.

Table of Contents

Sharper Quantitative Bounds

The upper bound on the differentially private sample complexity of a class H has a double exponential dependence on its Littlestone dimension $Ldim(H)$

[Alo+19] proved a lower bound which depends on $\log(Ldim(H))$

Can every class H be privately learned with sample complexity $poly(VC(H), \log(Ldim(H)))$?

Study of global stability

This paper introduced global stability as a definition of stability, and showed its connection to differentially privacy

Its natural to wonder whether global stability has connection with other definitions of stability

For example, [CMY23] viewed stability as a variant of replicability, and created a variant called list replicability, which has an algorithmic equivalence to global stability

Extension to multi-class classification and regression

It is natural to ask whether this equivalence holds for multi-class classification and regression problem

[JKT20] showed Private Learnable $\Rightarrow$ Online Learnable. But it is hard to show Online Learnable $\Rightarrow$ Private Learnable

Supplementary 1

Proposition (Stable Learner $\Rightarrow$ generalize well)

Suppose A is realizable learner ($loss(A(S)) = 0$ for any realizable S). D is a realizable distribution s.t. A is (n, η) stable and $\underset{S \sim D^n}{\mathbb{P}}[A(S) = h] \geq \eta$ for a hypothesis h, then

$$loss_D(h) \leq \frac{\ln(1/\eta)}{n}$$

Sketch Proof:

1) let $\alpha = loss_D(h)$. Since A is realizable, then A(S) = h means that $\exists$ h s.t. h is realizable

2) $\mathbb{P}$[h is realizable] $= (1 - \alpha)^n$. The probability that h correctly classifies the sample is equal to $1 - \alpha$. All n independent samples so $(1 - \alpha)^n$

3) $\underset{S \sim D^n}{\mathbb{P}}[A(S) = h] \Rightarrow \mathbb{P}$[h is realizable], so $\eta \leq (1 - \alpha)^n$. With the inequality $1 - \alpha \leq e^{-\alpha} \Rightarrow \eta \leq e^{-n\alpha} \Rightarrow \alpha \leq \frac{1}{n}\ln(\frac{1}{\eta})$