Sample-efficient proper PAC learning with approximate differential privacy

Badih Ghazi[1] **Noah Golowich**[2] Ravi Kumar[1] Pasin Manurangsi[1]

[1]Google Research

[2]MIT; work done while interning at Google Research

Overview: privacy-preserving PAC learning

- Machine learning models often trained on sensitive data; important to protect privacy of users' data

Source: cio.com

Source: sisense.com

- Our focus: fundamental **private PAC model** *[Kasiviswanathan et al., '08]*
- Recent development: connection between *private learnability* and *online learnability* *[Alon-Livni-Malliaris-Moran '19] [Bun-Livni-Moran '20]*
  - **This talk**: answer two open questions on "online learnability ⇒ private learnability" from *[Bun-Livni-Moran '20]*

Overview

1. **Background on Private PAC learning**
2. Sample-efficient proper private PAC learning
   - Key ingredient: irreducibility
3. Implications for sanitization.

Background: differential privacy

- Collection of individuals, each produces **example** $(x_i, y_i) \in X \times \{-1,1\}$
- **Dataset** $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, (randomized) learner A:

({smoker, 51 years, female}, cancer)

({smoker, 24 years, male}, no cancer)

({non-smoker, 32 years, female}, no cancer)

Private learning algorithm A

Classifier $f : X \to \{-1,1\}$ predicting whether cancer present.

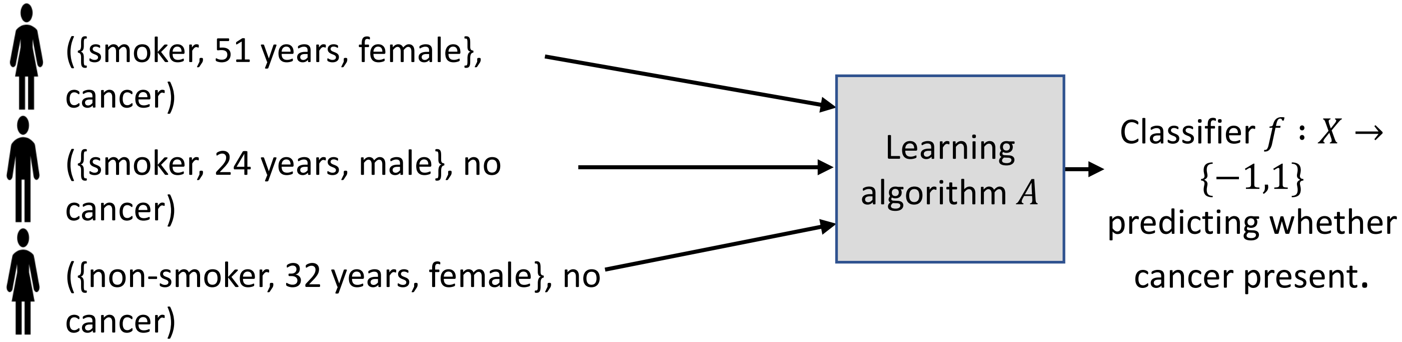
Definition: Algorithm A is **(ϵ, δ)-differentially private (DP)** if for all events E, for all *neighboring datasets* $S_n, S_n{}'$,

$$\Pr_A[A(S_n) \in E] \leq e^{\epsilon} \cdot \Pr_A[A(S_n') \in E] + \delta$$

Neighboring datasets: those which differ in a single example (x_i, y_i)

In this talk: $\epsilon \leq O(1)$ (e.g., $\epsilon = 0.01$), $\delta < 1/n^{\omega(1)}$ (e.g., $\delta = n^{-\log n}$)

PAC learning



- Given a *known* class F of **hypotheses**, i.e., functions $f : X \to \{-1,1\}$
- $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ is drawn i.i.d. from *unknown* distribution P on $X \times \{-1,1\}$
- Goal: algorithm $A(S_n)$ outputs $\hat{f}: X \to \{-1,1\}$ minimizing

$$\mathrm{err}_P(\hat{f}) := \Pr_{(x,y)\sim P}[\hat{f}(x) \neq y]$$

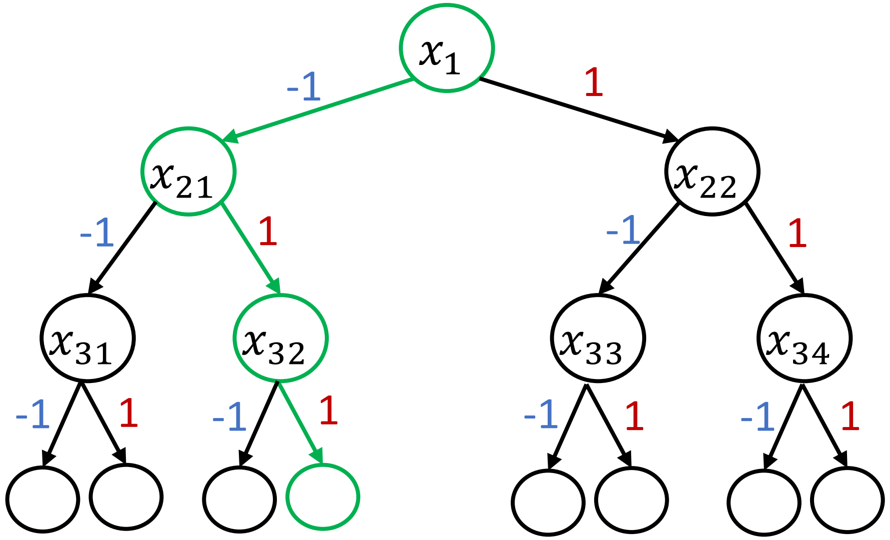
- In this talk: **realizable setting** (WLOG by *[Alon-Beimel-Moran-Stemmer, '20]*):

$$\boxed{\text{exists } f^* \in F \text{ so that } f^*(x) = y \text{ for all } (x, y) \in \mathrm{support}(P)}$$

- A is **proper** if $\hat{f} \in F$ almost surely, otherwise is **improper**

Background: private PAC learning, Littlestone dimension

- **Private PAC model**: algorithm A mapping $S_n \mapsto \hat{f}$ must be (ϵ, δ)-DP
- Hypotheses classes F with a private PAC learning algorithm achieving error $o(1)$ are exactly those with finite **Littlestone dimension** *[Alon-Livni-Malliaris-Moran '19] [Bun-Livni-Moran '20]*



Defn: For a binary tree with all internal nodes labeled by elements of X, edges labeled by $\{-1,1\}$:

- It is **shattered** by F if for each leaf ℓ there is some $f_\ell \in F$ which labels all nodes on the root-to-leaf path for ℓ according to the labels on the edges.
- E.g., for the green leaf: need $f_\ell(x_1) = -1, f_\ell(x_{21}) = 1, f_\ell(x_{32}) = 1$.

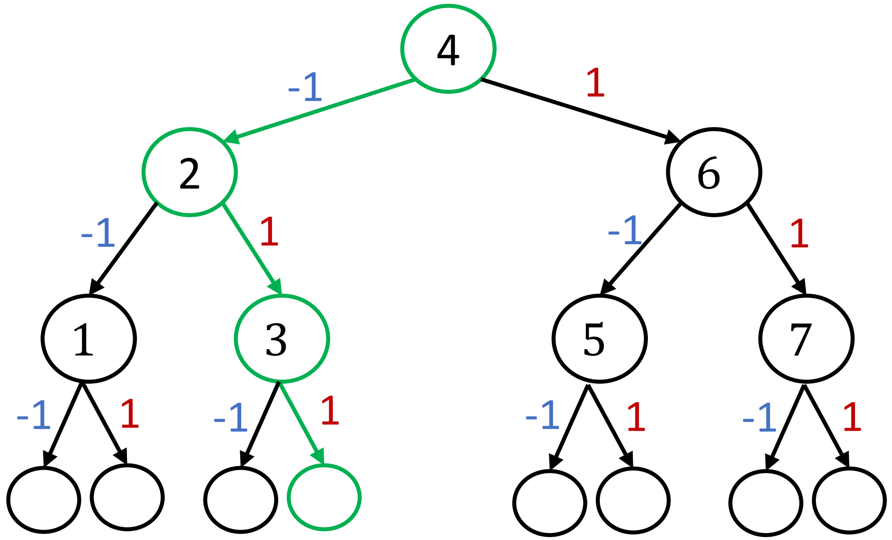
Defn: **Littlestone dimension** of hypothesis class F, denoted $\mathrm{Ldim}(F)$, is largest d so that there exists tree of depth d shattered by F.

- Finiteness of Littlestone dim. of F also characterizes its **online learnability**

Examples: finite Littlestone dimension classes

- Any finite class F has Littlestone dimension $\mathrm{Ldim}(F) \leq \log(|F|)$
- Class of threshold functions on $X = \{1, 2, \ldots, 2^d\}$ has $\mathrm{Ldim}(F) = d$
  - 2^d such thresholds; threshold i evaluates to 1 on $j \in X$ iff $i \leq j$

Example of shattered tree for $d = 3$:



Green leaf corresponds to threshold which evaluates to 1 on x iff $x \leq 3$

- **Throughout this talk: will use $\boldsymbol{d}$ to denote Ldim$(\boldsymbol{F})$**

Prior work: sample complexity of private & non-private learning

- Minimum number of samples n to achieve error $\alpha = o(1)$ in the:

(Non-private) PAC setting:

$$\boldsymbol{\Theta_{\alpha}(\mathrm{VCdim}(F))}$$

(where $VCdim(F)$ is the VC dimension of F) [Vapnik-Chervonenkis, '71]

Private PAC setting:

$$\boldsymbol{n \leq O_{\alpha,\epsilon,\delta}(2^{\mathrm{Ldim}(F)})}$$
$$\boldsymbol{n \geq \Omega(\log^{*}(\mathrm{Ldim}(F)))}$$

[Alon-Livni-Malliaris-Moran '19] [Bun-Livni-Moran '20]

Remarks:

- $\mathrm{VCdim}(F) \leq \mathrm{Ldim}(F)$ for all F; moreover, gap between them can be arbitrarily big.
- For private PAC learning, can't hope for bound *sublinear* in $\mathrm{Ldim}(F)$ if you want bound to depend only on $\mathrm{Ldim}(F)$ since there is F with $\mathrm{VCdim}(F) = \mathrm{Ldim}(F)$.

Overview

1. Background on Private PAC learning
2. **Sample-efficient proper private PAC learning**
    - **Key ingredient: irreducibility**
3. Implications for sanitization.

Sample-efficient proper private learning

- Let F be a hypothesis class of Littlestone dimension d, consisting of $f : X \to \{-1,1\}$
- Let P be a realizable distribution on $X \times \{-1,1\}$

Theorem: For $n = \tilde{O}(\frac{d^6}{\epsilon\alpha^2})$, there is an algorithm A which takes as input n i.i.d. samples from P, is (ϵ, δ)-DP, and outputs with high probability a hypothesis $\hat{f} \in F$ with classification error under P at most α (i.e., $\mathrm{err}_P(f) \leq \alpha$).

- Recall: that $\hat{f} \in F$ means A is **proper**
- *[Bun-Livni-Moran, '20]* showed a sample complexity bound of $n \approx \frac{2^{O(d)}}{\epsilon\alpha}$, and their learner *was not proper*

Proof overview: irreducibility

1. Show existence of an *improper* learner with polynomial sample complexity
   - Outputs **SOA classifier** for subclass satisfying special property: k-irreducibility
2. Use irreducibility and a min-max swap (i.e., *Sion's minimax theorem*) to "upgrade" the improper learner to a *proper* one

Definition: A hypothesis class G consisting of $f: X \to \{-1,1\}$ is **1-irreducible** if for any $x \in X$, there is some $b \in \{-1,1\}$ so that

$$\mathrm{Ldim}(\{g \in G : g(x) = b\}) = \mathrm{Ldim}(G)$$

For $k \geq 1$, ***k*-irreducibility** generalizes 1-irreducibility.

- Main idea: the **SOA classifier** for irreducible classes has certain "stability" properties conducive to the SOA classifier being private

SOA hypotheses & irreducibility

"**restriction** of G to (x, b)"

- For $G \subset F, b \in \{-1,1\}$: define $G|_{(x,b)} := \{g \in G : g(x) = b\}$
- For $G \subset F$, define SOA hypothesis $\mathrm{SOA}_G : X \to \{-1,1\}$, by:

$$\mathrm{SOA}_G(x) = \begin{cases} 1 & \text{if } \mathrm{Ldim}\left(G|_{(x,1)}\right) \geq \mathrm{Ldim}\left(G|_{(x,-1)}\right) \\ -1 & \text{otherwise} \end{cases}$$

- Example: point functions G on $X = \{x_1, \dots, x_5\}$; $G = \{g_1, \dots, g_5\}$:

X-value	g_1	g_2	g_3	g_4	g_5	SOA_G
x_1	1	-1	-1	-1	-1	-1
x_2	-1	1	-1	-1	-1	-1
x_3	-1	-1	1	-1	-1	-1
x_4	-1	-1	-1	1	-1	-1
x_5	-1	-1	-1	-1	-1	-1

G is **1-irreducible** if for any $x \in X$, there is some $b \in \{-1,1\}$ so that $\mathrm{Ldim}\left(G|_{(x,b)}\right) = \mathrm{Ldim}(G)$

- G is irreducible: $\mathrm{Ldim}(G) = 1$, and $\mathrm{Ldim}\left(G|_{(x,-1)}\right) = 1$ for all $x \in X$
  - Since $\mathrm{Ldim}\left(G|_{(x,1)}\right) = 0$ for all x, $\mathrm{SOA}_G(x) = -1$ for all $x \in X$

Simple properties of irreducibility

Lemma 1 (alternative phrasing of irreducibility defn): Suppose H is 1-irreducible. For $x \in X$ and $b \in \{-1,1\}$, $b = \mathrm{SOA}_H(x)$ if and only if $\mathrm{Ldim}\left(H|_{(x,b)}\right) = \mathrm{Ldim}(H)$.

Lemma 2 ("stability of SOAs"): Suppose that $H \subset G$, $\mathrm{Ldim}(H) = \mathrm{Ldim}(G)$, and that H is 1-irreducible. Then $\mathrm{SOA}_G = \mathrm{SOA}_H$, i.e., for all $x \in X$, $\mathrm{SOA}_G(x) = \mathrm{SOA}_H(x)$.

Proof is simple: fix any $x \in X$, suppose $\mathrm{SOA}_H(x) = 1$ (-1 is similar). Then:

$$\mathrm{Ldim}\left(G|_{(x,1)}\right) \geq \mathrm{Ldim}\left(H|_{(x,1)}\right) \underbrace{=}_{\text{Lemma 1}} \mathrm{Ldim}(H) = \mathrm{Ldim}(G)$$

and so $\mathrm{Ldim}\left(G|_{(x,1)}\right) = \mathrm{Ldim}(G)$, i.e., $\mathrm{SOA}_G(x) = 1 = \mathrm{SOA}_H\ (x)$.

"**restriction** of G to (x, b)"

Proof: SOA hypotheses

- For $G \subset F, b \in \{-1,1\}$: define $G|_{(x,b)} := \{g \in G : g(x) = b\}$.
- For $G \subset F$, define SOA hypothesis $\mathrm{SOA}_G : X \to \{-1,1\}$, by:

$$\mathrm{SOA}_G(x) = \begin{cases} 1 & \text{if } \mathrm{Ldim}\left(G|_{(x,1)}\right) \geq \mathrm{Ldim}\left(G|_{(x,-1)}\right) \\ -1 & \text{otherwise} \end{cases}$$

- Note: if G is 1-irreducible, never have $\mathrm{Ldim}\left(G|_{(x,1)}\right) = \mathrm{Ldim}\left(G|_{(x,-1)}\right)$.
- Main step of proof:

Depending only on P, **not** on the dataset S_n.

Lemma (relaxed global stability): Given P, there is a hypothesis $\sigma^* : X \to \{-1,1\}$ so that given a dataset $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn iid from P with $n = \mathrm{poly}(d)$, we can construct from S_n subclasses $\hat{G}_1, \ldots, \hat{G}_J \subset F$ so that:

1. Each $\mathrm{SOA}_{\hat{G}_j}$ has low population error w.h.p. (i.e., $\mathrm{err}_P(\mathrm{SOA}_{\hat{G}_j})$ is small)
2. With probability $\approx \frac{1}{d}$ over S_n, there is some $j \leq J$ so that $\mathrm{SOA}_{\hat{G}_j} = \sigma^*$.

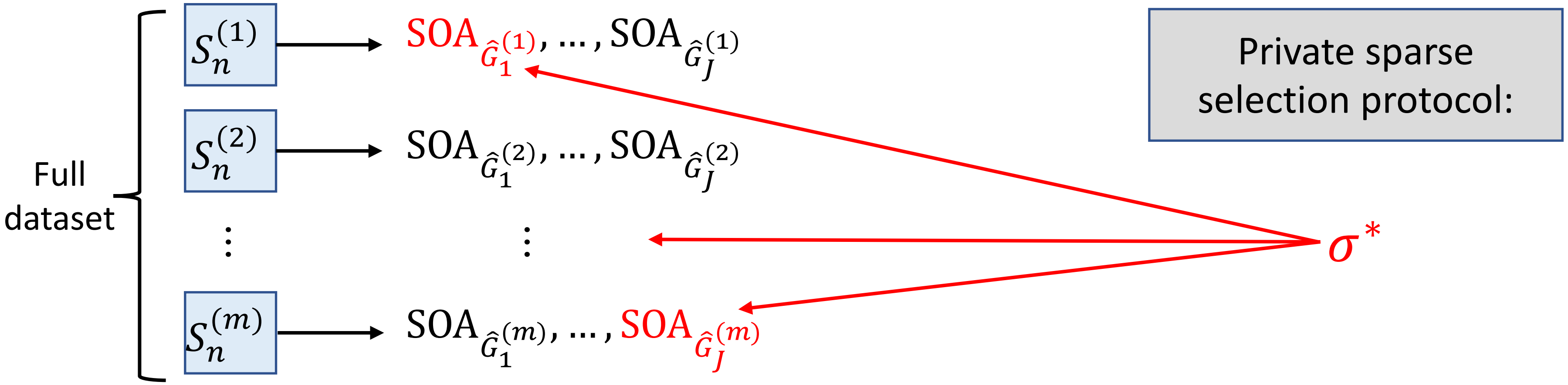
Using relaxed global stability

Lemma (relaxed global stability): Given P, there is a "special" hypothesis $\sigma^* : X \to \{-1,1\}$ so that given a dataset $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ drawn iid from P with $n = \mathrm{poly}(d)$, we can construct from S_n subclasses $\hat{G}_1, \dots, \hat{G}_J \subset F$ so that:

1. Each $\mathrm{SOA}_{\hat{G}_j}$ has low population error w.h.p. (i.e., $\mathrm{err}_P(\mathrm{SOA}_{\hat{G}_j})$ is small)
2. With probability $\approx \frac{1}{d}$ over S_n, there is some $j \le J$ so that $\mathrm{SOA}_{\hat{G}_j} = \sigma^*$.

Will have $J = 2^{O(d^2)}$

- **Consequence**: with $m \approx \tilde{O}(d)$ independent draws of S_n, can w.h.p discover some such σ^* -- turns out to be enough for private learnability (intuitively clear):
    - In particular, use a private **sparse selection protocol** *([BNS, '16; GKM, '20])*



Proof of "relaxed global stability" lemma

Lemma (relaxed global stability): Given P, there is a hypothesis $\sigma^* : X \to \{-1,1\}$ so that given a dataset $S_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ drawn iid from P with $n = \mathrm{poly}(d)$, we can construct from S_n subclasses $\hat{G}_1, \dots, \hat{G}_J \subset F$ so that:

1. Each $\mathrm{SOA}_{\hat{G}_j}$ has low population error (i.e., $\mathrm{err}_P(\mathrm{SOA}_{\hat{G}_j})$ is small)
2. With probability $\approx \frac{1}{d}$ over S_n, there is some $j \le J$ so that $\mathrm{SOA}_{\hat{G}_j} = \sigma^*$.

$\hat{P}_n$ is empirical distr. for S_n, i.e., uniform distribution on $\{(x_1, y_1), \dots, (x_n, y_n)\}$.

- Notation: for distribution Q and $\alpha > 0$, set $F_{Q,\alpha} := \{f \in F : \mathrm{err}_Q(f) \le \alpha\}$.
- **Idea**: condition on whether below assumption holds, where $\alpha > 0$ is some small parameter representing "acceptable" population error and $\alpha_\Delta \ll \alpha$:

Assumption: For a given sample S_n, it holds that $\mathrm{Ldim}(F_{\hat{P}_n,\alpha}) = \mathrm{Ldim}(F_{\hat{P}_n,\alpha-\alpha_\Delta})$ and $F_{\hat{P}_n,\alpha-\alpha_\Delta}$ is 1-irreducible.

Set $\sigma^* = \mathrm{SOA}_{F_{P,\alpha-\alpha_\Delta/2}}$

- If Assumption holds: by VC theory, $F_{\hat{P}_n,\alpha-\alpha_\Delta} \subset F_{P,\alpha-\alpha_\Delta/2} \subset F_{\hat{P}_n,\alpha}$, and so all 3 have equal Ldim; using irreducibility, by Lemma on prev. slide, $\mathrm{SOA}_{F_{P,\alpha-\alpha_\Delta/2}} = \mathrm{SOA}_{F_{\hat{P}_n,\alpha-\alpha_\Delta}}$.
- Else: find x so that $\mathrm{Ldim}(F_{\hat{P}_n,\alpha-\alpha_\Delta}|_{(x,1)}), \mathrm{Ldim}(F_{\hat{P}_n,\alpha-\alpha_\Delta}|_{(x,-1)}) < \mathrm{Ldim}(F_{\hat{P}_n,\alpha-\alpha_\Delta})$, "recurse" on $F|_{(x,1)}$ and $F|_{(x,-1)}$.

Generalization of 1-irreducibility

Definition: A hypothesis class G consisting of $f: X \to \{-1,1\}$ is ***k*-irreducible** if for any depth-k tree $\boldsymbol{x}$, there is some $b_1, \ldots, b_k \in \{-1,1\}$ so that

$$\mathrm{Ldim}\left(F|_{(\boldsymbol{x}_1,b_1),(\boldsymbol{x}_2(b_1),b_2),\ldots,(\boldsymbol{x}_k(b_{1:k-1}),b_k)}\right) = \mathrm{Ldim}(F)$$

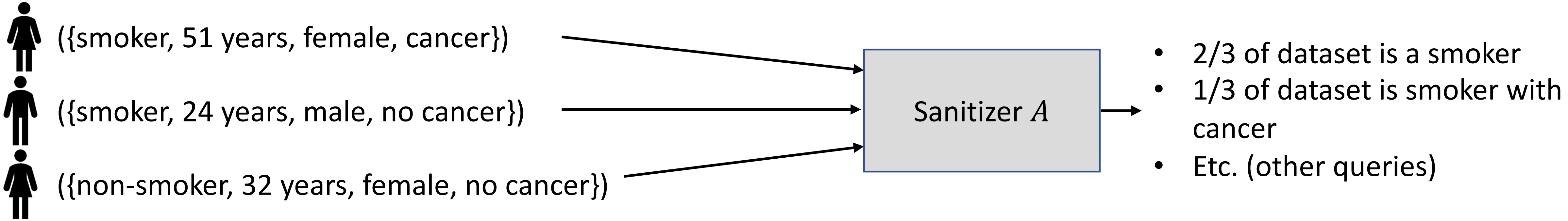
- In words: there is some leaf of the tree $\boldsymbol{x}$ so that the Ldim of the class restricted to that leaf is equal to the Ldim of F.
- Important for the general inductive step of the proof.

Overview

1. Background on Private PAC learning
2. Sample-efficient proper private PAC learning
    - Key ingredient: irreducibility
3. **Implications for sanitization.**

Background: sanitization *[Blum-Ligett-Roth, '08; Beimel-Nissim-Stemmer, '14]*

- **Sanitization** (i.e., **private query release**): give an estimate for the mean of each binary hypothesis $f \in F$ over a given dataset S.



Definition: Fix X, F. Algorithm A is a **sanitizer** for F with accuracy α and sample complexity n if it is (ϵ, δ)-DP and for all datasets $S = (x_1, \ldots, x_n) \in X^n$, $A(S)$ outputs Est: $F \to [-1,1]$, so that, with high probability, for all $f \in F$,

$$\left| \mathrm{Est}(f) - \frac{1}{n}\sum_{i=1}^{n} f(x_i) \right| \leq \alpha$$

Implications for sanitization

- *[Bousquet-Livni-Moran '20]:* "Private proper learning" ⇒ "sanitization"
- Our result: "Finite Littlestone dim." ⇒ "Private proper learning"; so:

Corollary: Suppose F has Littlestone dimension d & **dual Littlestone dimension** d^*.
For $n = \tilde{O}(\frac{d^6\sqrt{d^*}}{\epsilon\alpha^3})$, F has a sanitizer with sample complexity n and accuracy α.

- Dual Littlestone dimension d^* of F is the Littlestone dimension of the dual class of F
- Known that $d^* \leq 2^{2^{d+2}}$, and so also using *[Bun-Nissim-Stemmer-Vadhan, '15]:*

Corollary: F is sanitizable (i.e., has a sanitizer with sample complexity $\mathrm{poly}(1/\alpha)$) if and only if it has finite Littlestone dimension.

Open Questions

Thank you for listening!

- Main question: characterization of sample complexity of (proper & improper) learning with approximate DP, up to a constant (ideally)
  - VC dimension gives characterization for (non-private) PAC learning *[Vapnik, ‘98]*
  - Littlestone dimension does so for online learning *[Littlestone, ‘87; Ben-David, Pál-Shalev-Shwartz, ‘09]*
  - One-way public coin CC does so for PAC learning with pure DP *[Beimel-Nissim-Stemmer, ‘19; Feldman-Xiao, ‘14]*
- Intermediate questions:
  - Can we get $O(\mathrm{Ldim}(F))$ samples? (Can’t do better for F s.t. $\mathrm{Ldim}(F) = \mathrm{VCdim}(F)$)
  - Best known lower bound is $\Omega(\mathrm{VCdim}(F) + \log^* \mathrm{Ldim}(F))$ *[Alon-Livni-Malliaris-Moran, ‘20];* so can we get upper bound of $\mathrm{poly}(\mathrm{VCdim}(F), \log^* \mathrm{Ldim}(F))$?
- Can the sample complexity of proper private learning (w/ approximate DP) be asymptotically larger than that for improper private learning?
  - Answer is “yes” for pure DP *[Beimel-Brenner-Kasiviswanathan-Nissim, ‘14]*