Sample complexity bounds for differentially private learning

Kamalika Chaudhuri
University of California, San Diego

Daniel Hsu
Microsoft Research

Outline

1. Learning and privacy model
2. Our results: sample complexity bounds for differentially-private learning
3. Recap & future work

Part 1. Learning and privacy model

Data analytics with sensitive information

eCommerce: customers' browsing & purchase histories
Clinical studies: patients' medical records & test results
Genomic studies: subjects' genetic sequences

Patient 1

age	34
test #1	1.76
test #2	86.6
has flu?	1

Patient 2

age	31
test #1	1.62
test #2	67.5
has flu?	0

⋮

Learn something useful about whole population from data about individuals.

Data analytics with sensitive information

eCommerce: customers' browsing & purchase histories
Clinical studies: patients' medical records & test results
Genomic studies: subjects' genetic sequences

Patient 1

age	34
test #1	1.76
test #2	86.6
has flu?	1

Patient 2

age	31
test #1	1.62
test #2	67.5
has flu?	0

⋮

Learn something useful about whole population from data about individuals.

This work: learning a binary classifier from labeled examples, where each training example is an individual's sensitive information.

Data analytics with sensitive information

Data analytics with sensitive information

Q: If a classifier is learned from some individuals' sensitive data, can releasing / deploying the classifier in public violate the privacy of individuals from the training data?

Data analytics with sensitive information

Q: If a classifier is learned from some individuals' sensitive data, can releasing / deploying the classifier in public violate the privacy of individuals from the training data?

A: Yes! Even after standard "anonymization", and even when just releasing aggregate statistics, because an adversary could have side-information.

Example: genome-wide association studies

Has disease

1.00										
.190	1.00									
.216	.251	1.00								
.186	.117	.047	1.00							
.154	.011	.170	.083	1.00						
.190	.140	.102	.095	.139	1.00					
.270	.215	.294	.248	.140	.141	1.00				
.101	.085	.170	.056	.234	.099	.175	1.00			
.239	.071	.163	.111	.161	.093	.199	.157	1.00		
.471	.117	.243	.094	.144	.123	.283	.216	.274	1.00	
.179	.202	.132	.094	.087	.159	.207	.108	.092	.294	1.00

Correlations statistics

Healthy

1.00										
.141	1.00									
.099	.175	1.00								
.093	.199	.157	1.00							
.123	.283	.216	.274	1.00						
.159	.207	.108	.092	.294	1.00					
.088	.152	.075	.163	.156	.220	1.00				
.046	.161	.092	.072	.157	.143	.147	1.00			
.078	.392	.122	.229	.160	.172	.145	.177	1.00		
.045	.155	.135	.139	.110	.048	.126	.104	.169	1.00	
.178	.135	.102	.258	.314	.165	.147	.158	.131	.074	1.00

Correlations statistics

Wang *et al* (2009): able to combine side-information and published correlation statistics to determine whether an individual from the study was in disease group or healthy group.

Privacy-preserving machine learning

<u>Goals</u>: learn an accurate classifier from sensitive data while also preserving the privacy of the data.

<u>This work</u>: how many labeled examples are needed to achieve both of these goals simultaneously?

Goal 1: Differential privacy

What kind of privacy guarantee can a good learning algorithm provide?

Differential privacy guarantee [Dwork *et al*, 2006]:
an individual's inclusion in the training data does not change (much) what an adversary could learn about that individual's sensitive information.

Goal 1: Differential privacy

(Definition from [Dwork, *et al* 2006], specialized to learning [Kasiviswanathan, *et al* 2008])

A learning algorithm $\mathcal{A}\colon (\mathcal{X} \times \{0,1\})^* \to \mathcal{H}$
is α-*differentially private* if:

For all training sets S and S' differing in at most one example,

$$\forall \mathcal{G} \subseteq \mathcal{H}, \quad \frac{\Pr_{\mathcal{A}}[\mathcal{A}(S) \in \mathcal{G}]}{\Pr_{\mathcal{A}}[\mathcal{A}(S') \in \mathcal{G}]} \leq e^{\alpha}.$$

- Probability is over internal randomness of the learning algorithm.
- Algorithm must behave similarly given similar training sets.
- Smaller $\alpha \in [0,1]$ corresponds to stronger guarantee.

Goal 2: Learning

Standard statistical learning guarantees:

If S is an i.i.d. sample from a distribution $\mathcal{P}$ over $\mathcal{X} \times \{0,1\}$, then $\mathcal{A}(S)$ returns a hypothesis $h \in \mathcal{H}$ such that w.p. $\geq 1-\delta$ (over random draw of S and randomness in $\mathcal{A}$)

$$\mathrm{err}_{\mathcal{P}}(h) \leq \min_{h' \in \mathcal{H}} \mathrm{err}_{\mathcal{P}}(h') + \epsilon$$

where $\mathrm{err}_{\mathcal{P}}(\tilde{h}) = \Pr_{(x,y)\sim\mathcal{P}}[\tilde{h}(x) \neq y]$.

What was known
(previous work)

- Sample complexity for finite hypothesis classes or VC classes over discrete data domains.

  [Kasiviswanathan *et al*, 2008], [Blum *et al*, 2008], [Beimel *et al*, 2010]

$$C \cdot \left(\frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2}\right) \cdot \left(\min\{\log|\mathcal{H}|,\ \mathrm{VC}_{\mathcal{H}} \log|\mathcal{X}|\} + \log\frac{1}{\delta}\right)$$

- Related problems: (synthetic) data set release.

What was known
(previous work)

- Sample complexity for finite hypothesis classes or VC classes over discrete data domains.

  [Kasiviswanathan *et al*, 2008], [Blum *et al*, 2008], [Beimel *et al*, 2010]

$$C \cdot \left(\frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2} \right) \cdot \left(\min\{\log|\mathcal{H}|,\ \mathrm{VC}_{\mathcal{H}} \log|\mathcal{X}|\} + \log\frac{1}{\delta} \right)$$

- Related problems: (synthetic) data set release.

What about infinite classes & continuous data domains?

Part 2. Sample complexity bounds for differentially-private learning

Our results

1. Some bad news: no distribution-independent sample complexity upper bound possible for differentially-private learning.
2. Some hope: differentially-private learning possible if
   a. learner allowed some prior-knowledge, or
   b. privacy requirement is relaxed.

1. No distribution-independent sample complexity upper bound

Let $\mathcal{H}$ be the class of threshold functions on the unit interval $[0,1]$, and pick any positive real number M.

For every α-differentially private algorithm $\mathcal{A}\colon ([0,1]\times\{0,1\})^* \to \mathcal{H}$, there is a distribution $\mathcal{P}$ (with full support) over $[0,1]\times\{0,1\}$ such that:

1. There exists a threshold $h^* \in \mathcal{H}$ with $\mathrm{err}_{\mathcal{P}}(h^*) = 0$.

2. If S is an i.i.d. sample of size $m \le M$ from $\mathcal{P}$, then

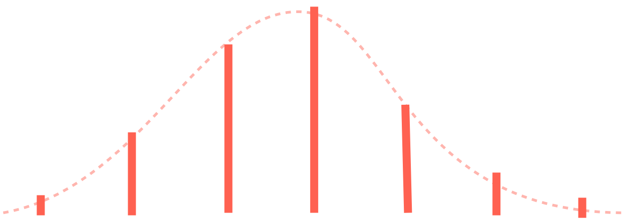
$$\Pr_{S\sim\mathcal{P}^m,\mathcal{A}}\left[\mathrm{err}_{\mathcal{P}}(\mathcal{A}(S)) > \frac{1}{5}\right] \ge \frac{1}{2}.$$

1. No distribution-independent sample complexity upper bound

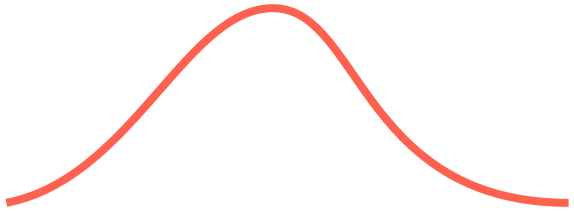
Implications:

1. No direct analogue of VC theorem for differentially-private learning.

2. Qualitative difference between finite hypothesis class / discrete data domains and infinite classes / continuous data domains.

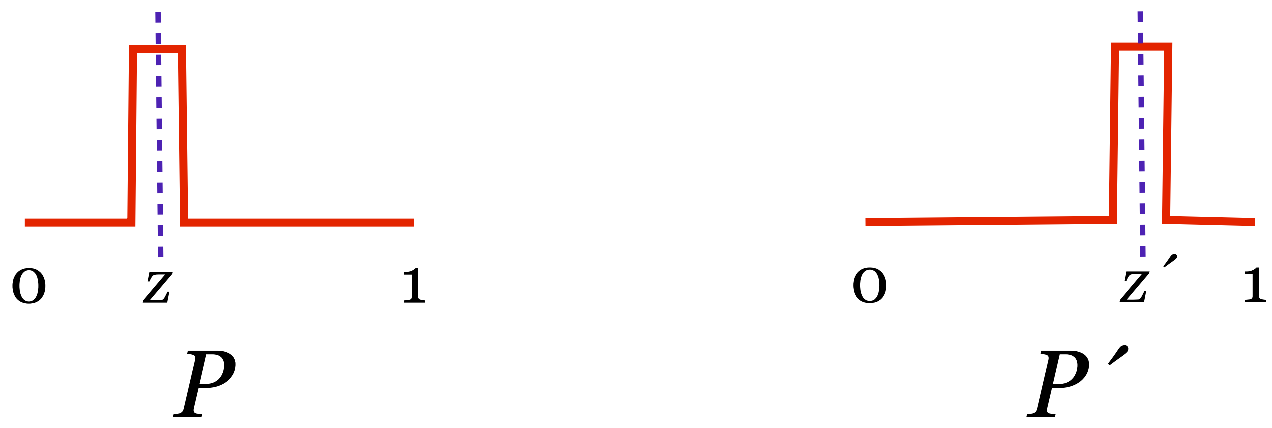


vs



1. No distribution-independent sample complexity upper bound

Proof idea: find data distributions P and P' such that a "successful" distribution over thresholds for P differs significantly from a "successful" distribution over thresholds for P'.



A differentially-private learner using just a small number of examples must behave similarly in both cases; therefore, it must fail for at least one of the cases.

2. Some hope for differentially-private learning

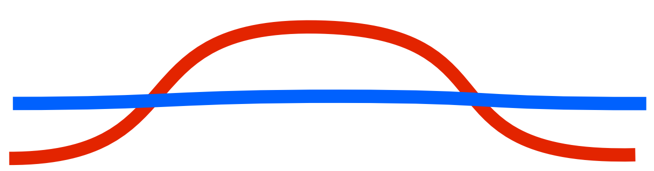
Possible ways around the lower-bound:

a. Allow learner access to prior-knowledge (or prior belief) about unlabeled data distribution.

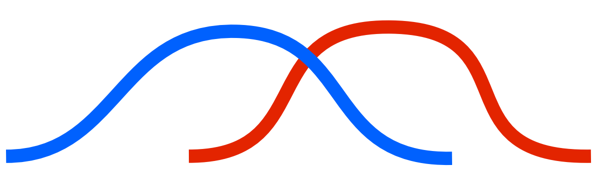
b. Only guarantee the differential privacy of the labels in the training data.

2(a). Upper bounds based on prior knowledge of unlabeled data distribution

- Allow learner access to a *reference distribution* *U* over unlabeled data *X*, chosen independently of the training data.
- Sample complexity upper bound depends on how close *U* is to *D* (true unlabeled data distribution).



U and *D* close



U and *D* far

2(a). Upper bounds based on prior knowledge of unlabeled data distribution

Let $\mathcal{P}$ be any distribution over $\mathcal{X} \times \{0,1\}$ with marginal $\mathcal{D}$ over $\mathcal{X}$. There is a constant $C > 0$ and an α-differentially private algorithm $\mathcal{A}_1$ s.t. given an i.i.d. sample S of size

$$|S| \geq C \cdot \left(\frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2} \right) \cdot \left(d_{\mathcal{U}} \cdot \log \frac{\kappa(\mathcal{U}, \mathcal{D})}{\epsilon} + \log \frac{1}{\delta} \right),$$

w.p. $\geq 1 - \delta$, $\mathcal{A}_1(S)$ returns a hypothesis $h \in \mathcal{H}$ with $\mathrm{err}_{\mathcal{P}}(h) \leq \min_{h' \in \mathcal{H}} \mathrm{err}_{\mathcal{P}}(h') + \epsilon$.

$d_{\mathcal{U}}$: doubling-dimension of disagreement metric w.r.t. $\mathcal{U}$.
$\kappa(\mathcal{U}, \mathcal{D})$: divergence measure between distributions $\mathcal{U}$ and $\mathcal{D}$.

2(a). Upper bounds based on prior knowledge of unlabeled data distribution

Let $\mathcal{P}$ be any distribution over $\mathcal{X} \times \{0,1\}$ with marginal $\mathcal{D}$ over $\mathcal{X}$. There is a constant $C > 0$ and an α-differentially private algorithm $\mathcal{A}_1$ s.t. given an i.i.d. sample S of size

$$|S| \geq C \cdot \left(\boxed{\frac{1}{\alpha\epsilon}} + \frac{1}{\epsilon^2} \right) \cdot \left(d_{\mathcal{U}} \cdot \log \frac{\kappa(\mathcal{U}, \mathcal{D})}{\epsilon} + \log \frac{1}{\delta} \right),$$

w.p. $\geq 1 - \delta$, $\mathcal{A}_1(S)$ returns a hypothesis $h \in \mathcal{H}$ with $\mathrm{err}_{\mathcal{P}}(h) \leq \min_{h' \in \mathcal{H}} \mathrm{err}_{\mathcal{P}}(h') + \epsilon$.

$d_{\mathcal{U}}$: doubling-dimension of disagreement metric w.r.t. $\mathcal{U}$.
$\kappa(\mathcal{U}, \mathcal{D})$: divergence measure between distributions $\mathcal{U}$ and $\mathcal{D}$.

2(a). Upper bounds based on prior knowledge of unlabeled data distribution

Let $\mathcal{P}$ be any distribution over $\mathcal{X} \times \{0,1\}$ with marginal $\mathcal{D}$ over $\mathcal{X}$. There is a constant $C > 0$ and an α-differentially private algorithm $\mathcal{A}_1$ s.t. given an i.i.d. sample S of size

$$|S| \geq C \cdot \left(\frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2}\right) \cdot \left(\boxed{d_{\mathcal{U}}} \cdot \log \frac{\kappa(\mathcal{U}, \mathcal{D})}{\epsilon} + \log \frac{1}{\delta}\right),$$

w.p. $\geq 1 - \delta$, $\mathcal{A}_1(S)$ returns a hypothesis $h \in \mathcal{H}$ with $\mathrm{err}_{\mathcal{P}}(h) \leq \min_{h' \in \mathcal{H}} \mathrm{err}_{\mathcal{P}}(h') + \epsilon$.

$d_{\mathcal{U}}$: doubling-dimension of disagreement metric w.r.t. $\mathcal{U}$.
$\kappa(\mathcal{U}, \mathcal{D})$: divergence measure between distributions $\mathcal{U}$ and $\mathcal{D}$.

2(a). Upper bounds based on prior knowledge of unlabeled data distribution

Let $\mathcal{P}$ be any distribution over $\mathcal{X} \times \{0,1\}$ with marginal $\mathcal{D}$ over $\mathcal{X}$. There is a constant $C > 0$ and an α-differentially private algorithm $\mathcal{A}_1$ s.t. given an i.i.d. sample S of size

$$|S| \geq C \cdot \left(\frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2} \right) \cdot \left(d_{\mathcal{U}} \cdot \log \frac{\boxed{\kappa(\mathcal{U}, \mathcal{D})}}{\epsilon} + \log \frac{1}{\delta} \right),$$

w.p. $\geq 1 - \delta$, $\mathcal{A}_1(S)$ returns a hypothesis $h \in \mathcal{H}$ with $\mathrm{err}_{\mathcal{P}}(h) \leq \min_{h' \in \mathcal{H}} \mathrm{err}_{\mathcal{P}}(h') + \epsilon$.

$d_{\mathcal{U}}$: doubling-dimension of disagreement metric w.r.t. $\mathcal{U}$.
$\kappa(\mathcal{U}, \mathcal{D})$: divergence measure between distributions $\mathcal{U}$ and $\mathcal{D}$.

2(a). Upper bounds based on prior knowledge of unlabeled data distribution

Example:

- H = n-dimensional linear separators through the origin
- U = uniform distribution on unit sphere (so $d_U = O(n)$)
- Unlabeled data distribution D close to uniform: $D(x) \leq c \cdot U(x)$
- Sample complexity upper bound:

$$C \cdot \left(\frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2} \right) \cdot \left(n \cdot \log \frac{c}{\epsilon} + \log \frac{1}{\delta} \right)$$

2(b). Label privacy

- Weaker privacy guarantee: only guarantee differential-privacy of the *labels*.
- Can still protect against some privacy attacks on training data.

A learning algorithm $\mathcal{A}\colon (\mathcal{X} \times \{0,1\})^* \to \mathcal{H}$ is α-*label private* if:

For all training sets $S, S' \subseteq \mathcal{X} \times \{0,1\}$ differing in at most one *label*,

$$\Pr_{\mathcal{A}}[\mathcal{A}(S) \in \mathcal{G}] \leq \Pr_{\mathcal{A}}[\mathcal{A}(S') \in \mathcal{G}] \cdot e^{\alpha} \quad (\forall \mathcal{G} \subseteq \mathcal{H})$$

2(b). Label privacy

- Label privacy avoids complications that arise with infinite hypothesis classes and continuous data domains
- Can obtain upper- and lower-bounds in terms of certain distribution-dependent complexity measures (covering number, doubling dimension).
- Bounds are (roughly) within $1/\alpha$ factor of non-private sample complexity bounds.

Recap & future work

1. Differential-privacy requirement rules out distribution-independent proper learning.
2. Some ways out:
   a. Data-dependent bounds based on prior-knowledge.
   b. Relaxed notion of privacy (label privacy).
3. Future directions:
   a. Improper learning (some work in discrete settings by [Beimel *et al*, 2010]).
   b. Other weaker notions of privacy.
   c. More general statistical estimation tasks.

Thanks!

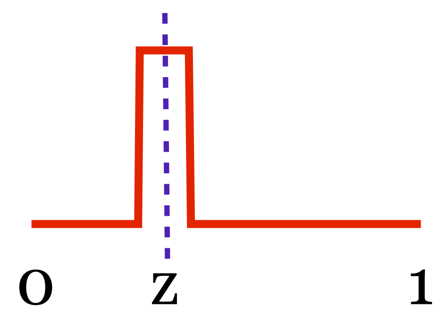
1. Bad news: no distribution-independent sample complexity upper bound

Idea: Consider a set of distributions $\{ P_z \}$ for $z \in [0,1]$: the marginal of each P_z over X is an even mixture of

(1) uniform on $[0,1]$, and

(2) uniform on $[z-\eta, z+\eta]$ (where $\eta = \Theta(\exp(-\alpha M))$);

and labels are given by threshold $h_z(x) = 1[x \geq z]$.



To show: Every α-differentially private learning algorithm using at most M training examples will fail on at least one distribution P_z.

2(a). Upper bounds based on *prior knowledge* of unlabeled data distribution

Example:

- H = n-dimensional linear separators through the origin
- U = uniform distribution on unit sphere (so $d_U = n$)
- Unlabeled data distribution D uniform outside $\Theta(1)$-width band around equator.
- Sample complexity upper bound:

$$C \cdot \left(\frac{1}{\alpha\epsilon} + \frac{1}{\epsilon^2}\right) \cdot \left(n^2 + n \cdot \log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)$$

Doubling dimension

- Hypothesis class *H* + unlabeled data distribution *D* ➙ disagreement metric space $(\mathcal{H}, \rho_{\mathcal{D}})$

$$\rho_{\mathcal{D}}(h, h') = \Pr_{x \sim \mathcal{D}}[h(x) \neq h'(x)]$$

- Doubling dimension is d if every ball of radius r can be covered by 2^d balls of radius $r/2$ (and no fewer).

- (Non-private) sample complexity bound due to Bshouty *et al* (2009) for noiseless setting:

$$C \cdot \frac{1}{\epsilon}\left(d + \log\frac{1}{\delta}\right)$$

Divergence κ(*U*,*D*)

$$\kappa(\mathcal{U},\mathcal{D}) = \inf\Big\{k > 0\colon \ \Pr_{x\sim\mathcal{D}}[x \in A] \le k \cdot \Pr_{x\sim\mathcal{U}}[x \in A] \quad \forall \text{ measurable } A\Big\}$$

(Quantifies absolute continuity of *D* w.r.t. *U*.)