Generalization and Learnability under Differential Privacy and its Variants

Yu-Xiang Wang

Based on joint works with Jing Lei and Steve Fienberg
Bottleneck of machine learning?

Privacy law

Food safety
Understand how brain works

Better treatment
Better education

Medical diagnosis
Clean energy
movie recommendation
Fraud detection

Increasing privacy awareness
The second Netflix Prize cancelled to settle a lawsuit
NYC’s Taxi data set breached

Vijay Pandurangan posts:

On Taxis and Rainbows
Lessons from NYC’s improperly anonymized taxi logs
Differentially private machine learning

- Data
  - Feature-label pairs
  - Unlabeled features
  - Feature points

- Learning Algorithm
  - Support Vector Machine
  - K-means clustering
  - Kernel density estimation

- Randomized algorithm
  - Classifier
    - K cluster centers
  - Estimated density function

- Bad Responses: Z Z Z
  - Ratio bounded

- Pr [response]
Example: Recommendation System

• Model based collaborative filtering.
  – Learning: \( f \leftarrow A(Z) \) uses all user data.
  – Prediction: \( y_i \leftarrow f(z_i) \) uses \( f \) and his own data.

• Setting:
  – Trust the service provider. Netflix is not an adversary.
  – Other users might be adversaries.

• If \( A \) is private, prediction is “post-processing”.
Synergy between learning and privacy

Underfitting
- $f$ is parsimonious
- Private information compressed.

Overfitting
- $f$ memorizes the dataset
- Knowing $f$ breaches privacy
Plan today

• Revisiting “What can be learned privately?”
  – Vapnik’s General Learning Setting
  – Characterizing private learnability

• To what extent can we weaken DP?
  – But still preserve the basic property that “privacy => generalization”.
  – A characterization of on-avg generalization.
Notations

• Data domain \( Z \) or \( \mathcal{X} \times \mathcal{Y} \)

• Hypothesis class \( \mathcal{H} \)

• Loss Function: \( \ell : \mathcal{H} \times Z \rightarrow \mathbb{R} \)

• Task: find \( h \in \mathcal{H} \) with low risk.
PAC Learning vs. General Learning Setting

(Agnostic) PAC Learning: Binary classification.
- Logistic Regression
  - Generalized Linear model
- Linear Regression
- RKHS Learning
  - Kernel SVM
- Multiclass classification
- Density estimation
- Stochastic convex optimization
  - K-means clustering
- Matrix factorization
  - Recommender system

General Learning Setting
- Reinforcement Learning
  - Online learning
Problems in general learning setting

<table>
<thead>
<tr>
<th>Problem</th>
<th>Hypothesis class $\mathcal{H}$</th>
<th>$\mathcal{Z}$ or $\mathcal{X} \times \mathcal{Y}$</th>
<th>Loss function $\ell$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAC Learning</td>
<td>$\mathcal{H} \subseteq {f : {0, 1}^d \rightarrow {0, 1}}$</td>
<td>${0, 1}^d \times {0, 1}$</td>
<td>$1(h(x) \neq y)$</td>
</tr>
<tr>
<td>Regression</td>
<td>$\mathcal{H} \subseteq {f : [0, 1]^d \rightarrow \mathbb{R}}$</td>
<td>$[0, 1]^d \times \mathbb{R}$</td>
<td>$</td>
</tr>
<tr>
<td>Density Estimation</td>
<td>Bounded distributions on $\mathcal{Z}$</td>
<td>$\mathcal{Z} \subseteq \mathbb{R}^d$</td>
<td>$- \log(h(z))$</td>
</tr>
<tr>
<td>K-means Clustering</td>
<td>${S \subseteq \mathbb{R}^d :</td>
<td>S</td>
<td>= k}$</td>
</tr>
<tr>
<td>RKHS classification</td>
<td>Bounded RKHS</td>
<td>RKHS x {0, 1}</td>
<td>$\max{0, 1 - y \langle x, h \rangle}$</td>
</tr>
<tr>
<td>RKHS regression</td>
<td>Bounded RKHS</td>
<td>RKHS x $\mathbb{R}$</td>
<td>$|\langle x, h \rangle - y|^2$</td>
</tr>
<tr>
<td>Sparse PCA</td>
<td>Rank-$r$ projection matrices</td>
<td>$\mathbb{R}^d$</td>
<td>$|h z - z|^2 + \lambda |h|_1$</td>
</tr>
<tr>
<td>Robust PCA</td>
<td>All subspaces in $\mathbb{R}^d$</td>
<td>$\mathbb{R}^d$</td>
<td>$|\mathcal{P}_h(z) - z|_1 + \lambda \text{rank}(h)$</td>
</tr>
<tr>
<td>Matrix Completion</td>
<td>All subspaces in $\mathbb{R}^d$</td>
<td>$\mathbb{R}^d \times {1, 0}^d$</td>
<td>$\min_{b \in h} |y \circ (b - x)|^2 + \lambda \text{rank}(h)$</td>
</tr>
<tr>
<td>Dictionary Learning</td>
<td>All dictionaries $\in \mathbb{R}^{d \times r}$</td>
<td>$\mathbb{R}^d$</td>
<td>$\min_{b \in \mathbb{R}^r} |hb - z|^2 + \lambda |b|_1$</td>
</tr>
<tr>
<td>Non-negative MF</td>
<td>All dictionaries $\in \mathbb{R}_+^{d \times r}$</td>
<td>$\mathbb{R}^d$</td>
<td>$\min_{b \in \mathbb{R}^r_+} |hb - z|^2$</td>
</tr>
<tr>
<td>Subspace Clustering</td>
<td>A set of $k$ rank-$r$ subspaces</td>
<td>$\mathbb{R}^d$</td>
<td>$\min_{b \in h} |\mathcal{P}_b(z) - z|^2$</td>
</tr>
<tr>
<td>Topic models (LDA)</td>
<td>$\mathbb{P}(\text{word}</td>
<td>\text{topic})$</td>
<td>Documents</td>
</tr>
</tbody>
</table>

An illustration of problems in the General Learning setting.
Learnability and Private Learnability

A learning algorithm: $A : \mathcal{Z}^n \rightarrow \mathcal{H}$ is consistent for distribution $\mathcal{D}$ if

$$\mathbb{E}_{Z \sim \mathcal{D}^n, z \sim \mathcal{D}} \mathbb{E}_{h \sim A(Z^n)} \ell(h, z) \rightarrow \min_{h \in \mathcal{H}} \mathbb{E}_z \ell(h, z).$$

Definition 1 (Learnability) A learning problem $(\mathcal{Z}, \mathcal{H}, \ell)$ is learnable if there exists an algorithm $A$ and rate $\xi(n)$, such that $A$ is consistent with rate $\xi(n)$ for any distribution $\mathcal{D}$ defined on $\mathcal{Z}$. [Note: this is agnostic, distribution-free learning!]

**Private Learnability**: Learnable by an $\epsilon$-DP algorithm, for an $\epsilon < \infty$
Defining Stability and AERM

- **Stability**: Any adjacent $Z$ and $Z'$, the difference in the expected risk $\rightarrow 0$ as $n \rightarrow \infty$.

- **AERM**: the estimate minimizes empirical risk as $n \rightarrow \infty$. 
What is known in non-private setting?

• PAC Learning (Binary Classification)
  – finite VC-dimension $\iff$ Learnability (BEHW–89)
  – Achieved by ERM.

• General Learning Setting
  – Strict Learnable by ERM $\iff$ Uniform Convergence (Vapnik–95)
  – $\exists$ a problem learnable, but ERM fails. (SSSS–10)
  – AERM + Stability $\iff$ Learnability (SSSS–10)
What is known in non-private setting?

Finite VC-Dimension

Fat-shattering dimension

Strictly Learnable problems: Uniform Convergence

AERM & Uniform Stability

General Learning Setting
What is known about private learnability?

- PAC Learning (on discrete domain):
  - SQ = Private SQ (BDMN-08)
  - PAC = Private PAC (KLNRS-08)
  - sample complexity on realizable setting (BNS-13).

- DP extensions of specific problems, or classes of problems.
  - (CMS-11, KST-12, BST-14) and many more.
What is known about private learnability?

PAC = PPAC
“What can we learn privately?”

SQ = PSQ. (‘‘SuLQ’’, Blum et. al. 05)

- Linear Regression
- Generalized Linear model
- Logistic Regression
- Stochastic convex optimization
- RKHS Learning
- K-means clustering
- Kernel SVM
- Matrix factorization
- Recommender system

- Multiclass classification
- Density estimation
- K-means clustering

General Learning Setting

- Reinforcement Learning
- Online learning
Our result

PAC = PPAC
(“What can be learned privately?”, Kasiviswanathan et. al., 08)
SQ=PSQ

Private Learnability = ∃ Private AERM

NOT Privately Learnable = ∄ Private AERM

General Learning Setting
Characterizing Private Learnability
Key ideas of the proof

Subsampling Lemma [BKN-13]:
If $A$ is $\epsilon$-DP on $Z$ of size $n$.
Then running $A$ on a random subsample of $Z$ with $\gamma n$ data points is $2\gamma \exp(\epsilon)$-DP.

Take $\gamma = 1/\sqrt{n}$
Key ideas of the proof

• Forward direction:
  – Private and AERM
  – Random subsample data (so that Privacy $\rightarrow 0$)
  – Privacy $\rightarrow$ Stability $\rightarrow$ Generalization (appeared in quite a few recent work, e.g., DFHPRR-14)

• Backward direction:
  – Given a private learning algorithm
  – Construct a new one by random subsampling
  – Show it’s AERM via distribution-free assumption.
Implications

• The task reduces to finding a private ERM

• A generic procedure that produces a learning algorithm for all privately learnable problems:

\[
\arg\min_{(A, \epsilon)} \left[ \epsilon + \sup_{Z \in \mathbb{Z}^n} \left( \mathbb{E}_{h \sim A(Z)} \hat{R}(h, Z) - \inf_{h \in \mathcal{H}} \hat{R}(h, Z) \right) \right]
\]

\(A : \mathbb{Z}^n \to \mathcal{H},\)  
\(A\) is \(\epsilon\)-DP
When can we learn but not privately learn?

Example in Chaudhuri and Hsu, 2011.
The difficulty of private classification in continuous domain (Chaudhuri and Hsu)

\[
\frac{p(A(Z))}{p(A(Z'))} \leq \exp(n\epsilon).
\]

Also, implicitly implied by sample complexity lower bound in [BKN13]
How to fix this?

• Lipschitz loss function, e.g., hinge loss.

• Drop the distribution-free requirement.
  – Private $\mathcal{D}$-learnability.
(ε,δ)-private learnability

• Extend subsampling lemma and stability lemma to (ε,δ)-DP.

• Results:
  – If we require δ = o(1/n),
  – Or if we require δ = o(1/poly(n)),
  – Approx. Private AERM = Approx. Private Learnability.
Are all learnable problems $(\varepsilon, \delta)$-privately learnable?

\[ \delta(n) \]

\[ O(e^{-n^2 \log n}) \]

Learnability = Approx. Private Learnability

Faster

Slower

Learnability $\neq$ Approx. Private Learnability
Story so far

• In general learning setting:
  – Private ERM learns all learnable problems.
  – Many problems are not privately learnable.
  – \((\varepsilon,\delta)\)-DP does not seem to solve the problem.

• Even if a problem is privately learnable...
  – might not be practical.
Practical frustrations with DP

• Need to add too much noise/ruin inferences.
  – Resulting in poor utility.
  – E.g., Contingency Table (Fienberg et. al. 2010), GWAS data (Yu et. al., 15), etc.

• Need a lot of tricks/hacks to work
  – E.g., “clipping” “rescaling” as in the Netflix data.

• Worst-case guarantee
  – Protects the worst possible data set.
  – Sensitive to outliers.
Same randomization, many interpretations

• How small needs $\epsilon$ be?

  – $\epsilon$-DP = 100
  – $\epsilon$-Personal DP for each person. $\epsilon<0.2$ for 95% of them.
  – On avg privacy: $\epsilon=0.1$

Weakening the privacy definition

• “A” outputs two distributions from Z and Z’.

• Any privacy definition should require the two distributions to be close.
  – $\epsilon$-DP $\Leftrightarrow$ $\epsilon$-Max-Divergence

• Use weaker distance measure?
Divergence privacy and f-divergence

• First seen in Barber & Duchi (2014).

\[ D_f(P \parallel Q) \equiv \int_{\Omega} f \left( \frac{dP}{dQ} \right) dQ. \]

• With P,Q being A(Z), A(Z’)

• When f = x \log x, this becomes **KL-divergence**.

\[ D_{KL}(P\parallel Q) = \int_{\Omega} \frac{dP}{dQ} \log \frac{dP}{dQ} dQ \]
On-Average KL-Privacy

• Differential Privacy:

\[
\sup_{Z, Z': d(Z, Z') \leq 1} \sup_{h \in \mathcal{H}} \log \frac{p_{h \sim A(Z)}(h)}{p_{h \sim A(Z')} (h)} \leq \epsilon
\]

• On-Average KL-Privacy:

\[
\mathbb{E}_{Z \sim \mathcal{D}^n, z \sim \mathcal{D}} \mathbb{E}_{h \sim A(Z)} \left[ \log \frac{p_{h \sim A(Z)}(h)}{p_{h \sim A([Z^-1, z])} (h)} \right] \leq \epsilon.
\]
On-Average KL-Privacy

• Measures the **average privacy loss** for a particular data generating distribution.

• Unaffected by rare pathological cases.

• Adapt to easy distributions.
Properties of on-average KL-Privacy

- Inherent properties of DP
  - Small group composition
  - Adaptive Composition (caveat:
    - Closed to post-processing
  - Does not need bounded loss function!
- When the loss function is bounded, the same algorithm guarantees DP.
Reusable Holdout/Adaptive Data Analysis

• A: learning algorithm output h.

\[ A \text{ is } \varepsilon\text{-DP } \Rightarrow \text{A has generalization error bound } \varepsilon \]

Definition:
Generalization error = \[ E |\text{Risk} - \text{Empirical Risk}|. \]


Characterizing the generalization

$$p(h) \propto \exp(-\mathcal{L}(Z))\pi(h)$$

**Theorem:** If $A$ is a posterior sampling algorithm for some model:

$$\varepsilon\text{-on-average KL-Privacy} \iff \varepsilon\text{-on-avg-generalizing}$$

For each distribution separately

**Definition:**

On-avg-generalization = $| \text{Risk} - \mathbb{E} \text{Empirical risk} |$

Why posterior sampling?

• A variational justification:
  – It arises out of a maximum entropy framework.

\[
\min_{\mathcal{A}} -H(\mathcal{A}(Z)|Z) + \mathbb{E}\mathcal{L}(\mathcal{A}(Z), Z)
\]

– The optimal solution:
\[
\mathcal{A}^*(Z) \sim p(h|Z) \propto \exp(-\mathcal{L}(h, Z))
\]

Why posterior sampling?

\[
\begin{align*}
\arg\min_{(A, \epsilon)} \left[ \epsilon + \sup_{Z \in \mathcal{Z}^n} \left( \mathbb{E}_{h \sim A(Z)} \hat{R}(h, Z) - \inf_{h \in \mathcal{H}} \hat{R}(h, Z) \right) \right] \\
\text{A is } \epsilon\text{-DP}
\end{align*}
\]

\[
\begin{align*}
\min_{A} -H(A(Z)|Z) + \mathbb{E} \mathcal{L}(A(Z), Z)
\end{align*}
\]
Why posterior sampling?

• A statistical justification:
  – Near optimal efficiency
  – Asymptotic normality
  – Works even under model misspecification.

Implication to adaptive data analysis

• Dwork et. al. 2015: Max-Information

\[ k\text{-max-information} \Rightarrow k/n\text{-on-average KL-Privacy} \]

For any distribution

• Russo & Zou 2015: Mutual information

\[ I(A(Z), Z) \leq \text{On-Avg-Gen.} \leq \sigma \sqrt{2I(A(Z), Z)} \]

For each distribution separately

The first lower bound of this form.
An example on Linear Regression

- Each user is \((x, y)\). We have 100 of them.
- Assume \((x, y)\) are inside \([-2, 2] \times [-1, 1]\)

- We want to privately fit a linear regression
  - \(y = x \beta_1 + \beta_0\)
  - From a bounded space \((\beta_1, \beta_0)\) in \([-2, 2]^2\)

- Loss function is \((y - x \beta_1 - \beta_0)^2\)

- Privately release through posterior sampling
  \[
  (\hat{\beta}_0, \hat{\beta}_1) \sim \frac{1}{K} e^{-\gamma \sum_{i=1}^{n} (y_i - \beta_0 - x_i \beta_1)^2}
  \]
Experiment on Linear Regression

Linear regression

- Est. Generalization
- KL-Privacy
- KL-Privacy/\gamma
- Excess risk
- Diff. Privacy loss $\epsilon$

4 orders of magnitude!

<<More privacy>>

$\gamma$

>>More utility>>
Experiment: normal mean

Normal mean estimation

- Est. Generalization
- KL-Privacy
- KL-Privacy/\gamma
- Excess risk
- Diff. Privacy loss $\epsilon$
Summary of On-Avg-KL Privacy

• No changes in algorithm
  – Average case privacy guarantee.
  – Practically meaningful
  – Esp. , when $\epsilon$ is too large.

• Characterizing the on-average generalization
  – Lower bounds of bias in terms of mutual information.
## Variants of Differential Privacy

<table>
<thead>
<tr>
<th>Privacy definition</th>
<th>$Z$</th>
<th>$z$</th>
<th>Distance (pseudo)metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pure DP</td>
<td>$\sup_{Z \in \mathcal{Z}^n}$</td>
<td>$\sup_{z \in \mathcal{Z}}$</td>
<td>$D_\infty(P \parallel Q)$</td>
</tr>
<tr>
<td>Approx-DP</td>
<td>$\sup_{Z \in \mathcal{Z}^n}$</td>
<td>$\sup_{z \in \mathcal{Z}}$</td>
<td>$D_\delta(P \parallel Q)$</td>
</tr>
<tr>
<td>Personal-DP</td>
<td>$\sup_{Z \in \mathcal{Z}^n}$</td>
<td>$\sup_{z \in \mathcal{Z}}$</td>
<td>$D_{\infty}(P \parallel Q)$ or $D_\delta(P \parallel Q)$</td>
</tr>
<tr>
<td>KL-Privacy</td>
<td>$\sup_{Z \in \mathcal{Z}^n}$</td>
<td>$\sup_{z \in \mathcal{Z}}$ for each $z$</td>
<td>$D_{KL}(P \parallel Q)$</td>
</tr>
<tr>
<td>TV-Privacy</td>
<td>$\sup_{Z \in \mathcal{Z}^n}$</td>
<td>$\sup_{z \in \mathcal{Z}}$</td>
<td>$|P - Q|_{TV}$</td>
</tr>
<tr>
<td>Rand-Privacy</td>
<td>$\sup_{Z \in \mathcal{Z}^n}$</td>
<td>$\sup_{z \in \mathcal{Z}}$</td>
<td>$D^{\delta_2}(P \parallel Q)$</td>
</tr>
<tr>
<td>On-Avg KL-Privacy</td>
<td>$1 - \delta_1$ any $\mathcal{D}^n$</td>
<td>$1 - \delta_1$ any $\mathcal{D}$</td>
<td>$D_\infty(P \parallel Q)$</td>
</tr>
</tbody>
</table>

For references of these privacy notions, please refer to the paper: http://arxiv.org/abs/1605.02277

Table 1. Summary of different privacy definitions.
Their connections to generalization

- KL-Privacy
- TV-Privacy
- On-Avg KL-Privacy
- For MaxEnt Sampling
- pinsker
- For insensitive queries
- Jensen
- If small variance
- Bounded
- Markov
- 1-δ Generalization

- Pure-DP
- Approx-DP
- Personal-DP
- Random-DP
- E Generalization
In summary

• Two recent work that investigates the connections of privacy and learning.

• Formalize the equivalence of generalization with some notion of privacy (for MaxEnt mechanisms).

• A practically useful interpretation of DP-algorithms.

• Towards practical privacy protection.